



Volume 6

Issue 1

2019

International Journal of  
Assessment Tools in Education

International Journal of  
Assessment Tools in Education

International Journal of  
Assessment Tools in Education

<http://ijate.net/>

e-ISSN: 2148-7456



e-ISSN 2148-7456

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

**Volume 6**

**Issue 1**

**2019**

**Dr. İzzet KARA**

Editor in Chief

International Journal of Assessment Tools in Education

Pamukkale University,

Education Faculty,

Department of Mathematic and Science Education,

20070, Denizli, Turkey

Phone : +90 258 296 1036

Fax : +90 258 296 1200

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

Publisher : İzzet KARA

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

Online ISSN: 2148-7456

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

<http://dergipark.gov.tr/ijate>

Design & Graphic: IJATE

### **Support Contact**

Dr. İzzet KARA

Journal Manager & Founding Editor

Phone : +90 258 296 1036

Fax : +90 258 296 1200

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

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

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



## International Journal of Assessment Tools in Education

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

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

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

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

**Editor in Chief**

Dr. Izzet Kara

**Editors**

Dr. Özen Yıldırım, *Pamukkale University*, Turkey

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

**Section Editor**

Dr. H.İbrahim Sari, *Kilis 7 Aralık University*, Turkey

**Editorial Board**

Dr. Hafsa Ahmed, *National University of Modern Languages*, Pakistan

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

Dr. Murat Balkıs, *Pamukkale University*, Turkey

Dr. Gülşah Başol, *Gaziosmanpaşa University*, Turkey

Dr. Bengü Börkan, *Boğaziçi University*, Turkey

Dr. Gülşah Başol, *Gaziosmanpaşa University*, Turkey

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

Dr. Gülşah Başol, *Gaziosmanpaşa University*, Turkey

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

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

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

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

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

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

Dr. Nuri Doğan, *Hacettepe University*, Turkey

Dr. Erdinç Duru, *Pamukkale University*, Turkey

Dr. Selahattin Gelbal, *Hacettepe University*, Turkey

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

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

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

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

Dr. Orhan Karamustafaoglu, *Amasya University*, Turkey

Dr. Yasemin Kaya, *Atatürk University*, Turkey

Dr. Hulya Kelecioğlu, *Hacettepe University*, Turkey

Dr. Hakan Koğar, *Akdeniz University*, Turkey

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

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

Dr. Ibrahim A. Njodi, *University of Maiduguri*, Nigeria

Dr. Jacinta A. Opara, *Kampala International University*, Uganda



Dr. Nesrin Ozturk, *Ege University*, Turkey

Dr. Turan Paker, *Pamukkale University*, Turkey

Dr. Abdurrahman Sahin, *Pamukkale University*, Turkey

Dr. Ragip Terzi, *Harran University*, Turkey

Dr. Hakan Türkmen, *Ege University*, Turkey

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

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

**English Language Editors**

Dr. Hatice Altun, *Pamukkale University*, Turkey

Dr. Çağla Atmaca, *Pamukkale University*, Turkey

Dr. Sibel Kahraman, *Pamukkale University*, Turkey

**Copy & Language Editor**

Anıl Kandemir, *Middle East Technical University*, Turkey

## Table of Contents

### *Research Article*

---

1. [Teachers' Test Construction Skills in Senior High Schools in Ghana: Document Analysis / Pages: 1-8](#)  
Frank Quansah, Isaac Amoako, Francis Ankomah
2. [An Effective Way to Provide Item Validity: Examining Student Response Processes / Pages: 9-24](#)  
Omer Kutlu, Hatice Cigdem Yavuz
3. [Can Factor Scores be Used Instead of Total Score and Ability Estimation? / Pages: 25-35](#)  
Abdullah Faruk Kılıç
4. [Impact of Emotional Literacy Training on Students' Emotional Intelligence Performance in Primary Schools / Pages: 36-47](#)  
Kerem Coskun, Yücel Oksuz
5. [An Investigation of Item Bias of English Test: The Case of 2016 Year Undergraduate Placement Exam in Turkey / Pages: 48-62](#)  
Rabia Akcan, Kübra Atalay Kabasakal
6. [The Post-Graduate Academic English Language Skills and the Language Skills Measured by the Iranian PhD Entrance Exam: A Test Reform and Curriculum Change / Pages: 63-79](#)  
Shiela Kheirzadeh, S. Susan Marandi, Mansoor Tavakoli
7. [Development of a Measurement Tool for Sustainable Development Awareness / Pages: 80-91](#)  
Ayşe Ceren Atmaca, Seyit Ahmet Kiray, Mustafa Pehlivan
8. [The Impact of Ignoring Multilevel Data Structure on the Estimation of Dichotomous Item Response Theory Models / Pages: 92-108](#)  
Hyung Rock Lee, Sunbok Lee, Jaeyun Sung
9. [Development of Exposure to English Scale and Investigation of Exposure Effect to Achievement / Pages: 109-124](#)  
Mustafa Gökcan, Derya Çobanoğlu Aktan
10. [Adaptation of Physics Metacognition Inventory to Turkish / Pages: 125-137](#)  
Zeynep Koyunlu Ünlü, İlbilge Dökme
11. [Performance Evaluation Using the Discrete Choquet Integral: Higher Education Sector / Pages: 138-153](#)  
Seher Nur Sülkü, Deniz Koçak
12. [Improved Performance of Model Fit Indices with Small Sample Sizes in Cognitive Diagnostic Models / Pages: 154-169](#)  
Hueying Tzou, Ya-Huei Yang

## Teachers' Test Construction Skills in Senior High Schools in Ghana: Document Analysis

Frank Quansah <sup>1,\*</sup>, Isaac Amoako<sup>1</sup>, Francis Ankomah<sup>1</sup>

<sup>1</sup> Department of Education and Psychology, University of Cape Coast, UCC, PMB, Cape Coast, Ghana.

### ARTICLE HISTORY

Received: 29 September 2018

Revised: 05 December 2018

Accepted: 11 December 2018

### KEYWORDS

Content relevance,  
Reliability,  
Validity,  
Representativeness,  
Test specification

**Abstract:** Assessment, specifically test construction, forms a critical part of the teaching and learning process. This aspect of teachers' responsibility has been questioned by several authorities in contemporary times. The study explored the test construction skills of Senior High Schools (SHS) teachers in the Cape Coast Metropolis. Using a qualitative document analysis, samples of End-of-Term Examination papers in Integrated Science, Core Mathematics and Social Studies in three selected SHS in the Cape Coast Metropolis were randomly (Lottery method) selected. The assessment tasks on the sampled instruments were critically examined by experts in the area of Educational Measurement and Evaluation. The results revealed that the teachers have limited skills in the construction of end-of-term examination. This was evident as issues were found with the content representativeness and relevance of the test, reliability, and fairness of the assessment tasks which were evaluated. It was recommended that head teachers should take up the challenge of inviting resource persons from recognised academic institutions to organise workshops for teachers on a regular basis to sharpen their skills on effective test construction practices.

## 1. INTRODUCTION

In the management of schools in Ghana, teachers, schools' management and policymakers in the course of or after teaching and sometimes before classroom teaching need to make decisions concerning teaching and learning. These decisions are made based on information gathered from the students' learning. Generally, this information gathering procedure denotes assessment. Nitko (2001) explained assessment as a process of obtaining information which is used for making decisions about students, curricula and programmes, and educational policy. Assessment, therefore, involves the utilisation of empirical data on students' learning to improve programmes and enhance students' learning (Allen & Yen, 2002). Scholars have pointed out that assessment systems have a significant effect on learning characteristics and personalities as children become young adults and then adults (Crooker & Algina, 2008;

---

CONTACT: Frank Quansah ✉ [fquansah99@gmail.com](mailto:fquansah99@gmail.com) 📧 Department of Education and Psychology, University of Cape Coast, UCC, PMB, Cape Coast, Ghana.

ISSN-e: 2148-7456 / © IJATE 2019

Ecclestone & Pryor, 2003; Nitko, 2001). Thus, the effects of assessment can continue through a learner's life of formal learning.

In the school setting, a test is generally used as an assessment tool for obtaining information about students' learning. It should be made clear at this point that testing is a key component in educational assessment. In testing what students know or have learnt in an area of study, well-crafted test items should be used. Tamakloe, Amedahe and Atta (1996) described a test as a device or procedure for measuring a sample of an individual's behaviour in a specific learned activity or discipline. Crooker and Algina (2008) further gave a description of test to be a standard procedure for obtaining a sample of behaviour from a specified domain. These tests are normally administered to students after a period of instruction, if for achievement purposes. Considering the sensitive role that information from a test play in making educational decisions for students as well as management, it is important to say that both test developers and users must make conscious effort to improve the validity and the reliability of the test in order to get objective information that approximate the individual's true characteristic, which the test developer seeks to estimate.

Unfortunately, test construction role of teachers has been reported as a main source of anxiety, especially with teachers with few years of teaching experience (Ebinye, 2001). This anxiety, according to Ebinye (2001), largely stems from inadequate test construction skills of these teachers. Scholars have also argued that test construction among teachers is not encouraging (e.g., Amedahe, 1989; Hamafyelto, Hamman-Tukur & Hamafyelto, 2015; Kazuko, 2010). The implication is that teachers may end up taking inaccurate information about student learning. For instance, Ololube (2008), which assessed the test construction skills of teachers in Nigeria, found poor test construction skills among non-professional teachers. Another study by Onyechere (2000) found that most teachers construct poor items which actually failed to function as it was supposed to. Some teachers, acknowledging that they have weak test construction skills resort to past or already existing questions to assess students (Onyechere, 2000). Teachers in the Borno State, Nigeria, were also found to construct items with lower levels of cognitive operations.

Similar findings have been found in Ghana. Amedahe (1989), in his study, found that SHS teachers in the Central Region of Ghana have inadequate skills in constructing both essay and objective type tests. According to the Curriculum, Research and Development Division [CRDD] of Ghana Education Service (GES) (1999), Junior High School teachers all over Ghana are found to have inadequate competencies in testing practices. Etsey's (2003) supported the views of CRDD (1999) and stated that the Division of Teacher Education of GES should authorize curriculum planners in education within the country to make assessment courses compulsory and as well prioritise these courses in the first 2 years in teacher training colleges.

In Ghana, Quansah and Amoako (2018) found that SHS teachers in the Cape Coast Metropolis have a negative attitude towards test construction. The authors specifically found a poor attitude of teachers in the planning of test, item writing, item review and assembling of the items. Quansah and Amoako concluded that this attitude of teachers had an effect on the quality of test used for assessing students. It is of essence to state that the poor attitude might not be due to their inadequate skills but also from the fact that some teachers see test construction as a burden. Exploring the test construction skills of teachers is significant if objective and accurate information are to be gathered from students in the teaching and learning process.

Moreover, previous studies employed self-reported means to describe teachers' skills in test construction. This measurement procedure does not appropriately estimate the skills of teachers in test construction. Majority of these studies gathered their information through administering questionnaires to the respondents or by interviewing them. The mere asking of questions about how these teachers construct test items do not provide a comprehensive view of the skills

teachers have. It is even likely that these teachers would provide responses which do not reflect their actual practice. In actual sense, previous studies just provide information about teachers' testing or test construction practices through the lens of the same teachers. It is essential to conduct an exploratory study to critically examine some questions crafted by these teachers to find out whether they have the competencies in test construction. These crafted questions serve as the "end-product" of their skills which is being put to use. This paper, therefore, explores the skills of SHS teachers in the Cape Coast Metropolis. The paper, particularly, assessed the content of the documents (samples of examination questions) with regards to five hypothetical dimensions: (a) content representativeness and relevance; (b) thinking processes and skills represented; (c) reliability and objectivity; (d) fairness to different students; and (e) practicality.

## **2. METHOD**

The research methodology for this study is qualitative document analysis. This study seeks to review and evaluate documents (Creswell, 2014). Just like other forms of the use of document analysis is to examine samples of previous examination papers in SHSs in the Cape Coast Metropolis in order to give meaning and understanding of teachers' test construction skills (Corbin & Strauss, 2008). Samples of End-of-Term Examination papers in Integrated Science, Core Mathematics and Social Studies in three selected SHS in the Cape Coast Metropolis were randomly (Lottery method) selected. These papers were used for summative assessment and thus, the questions were crafted by the classroom teachers in conjunction with the examination board of the school. This means that the papers went through some form of evaluation before they were administered. The question papers selected were papers between 2015 and 2018 in three subjects: Integrated Science, Core Mathematics, and Social Studies. Specifically, 5 samples of question papers were selected based on each subject from each school. In total, 15 samples of examination papers were taken from each of the three schools. In all, 45 samples of examination papers were sampled from the three schools.

The assessment tasks on the sampled instruments were critically examined by experts in the field of Measurement and Evaluation. The examination of the papers took four months. Before the questions were examined, the scheme of work of the various subjects selected was taken. There was also an interaction with the teachers on the areas which were covered for the term. We made a lot of effort to ensure the schools' anonymity, confidentiality and privacy in the data gathered. Consent of the teachers together with the examination committee was sought before the data was gathered. We employed the qualitative content analysis to analyse texts in the pre-defined dimensions.

### **2.1. Description of the Papers**

The Integrated Science papers were for Form 1 (first year/grade 10) students in the selected Senior High Schools. All the Integrated Science papers, the assessment tasks were in two sections: paper A and B. Paper A carried 40 points whereas the paper B carried 60 points. The paper A consisted of 40 multiple choice items with four options and students were required to respond to all the questions. The paper B was the essay section which also had two parts: Part I and II. The part I was a practical compulsory question which had five sub-sections. Part II of the paper B had four questions of which the students were required to answer only two. For all the Integrated Science papers, the examinees (students) were required to answer the question within a two-hour duration.

For the Core Mathematics papers, the samples were taken from Form 2 (second year/grade 11) students in the selected Senior High Schools. For all the Core Mathematics papers, the test comprised two sections (i.e., A & B). Section A comprised 40 multiple choice items and section B had seven essay type questions with their sub-questions where students were required to

answer five items. Question one of Section B was always a compulsory question, however, students were to choose other four) from the other six questions.

The Social Studies papers were also made up of two sections: Section A and Section B. The first sections comprised of 40 multiple-choice items which the students were required to use 50 minutes in responding to it. The second section had four essay questions which students were required to select three. Each of the essay questions carried 20 points. The paper was for Form 2 (second year/grade 11) students.

### **3. RESULTS AND DISCUSSION**

The results from the examination of the papers are captioned into the following sections: content representativeness and relevance; thinking processes and skills represented; reliability and objectivity; fairness to different students; and practicality.

#### **3.1. Content Representativeness and Relevance**

After careful evaluation of the test instruments, it was evident that the test developer who is the subject teacher failed to sample adequately to cover all the content areas listed in the scheme of work for the relevant term. Analysis from the papers revealed that the content of these papers focused on a few of the areas taught. It was evident that the items on the instruments (tests) did not adequately sample the content taught. This implies that a student who attains 90% (Distinction) cannot be addressed as having adequate mastery of the content taught since he/she was not assessed in all the areas taught. Likewise, a student who obtains 35% (Fail) cannot be referred to as lacking mastery of the content taught. This is because the higher scoring student might have specialized in the areas which were sampled while the lower scoring student did not. It is possible that the higher scoring student lacks mastery over the three content areas which assessment instruments did not cover. The result from these assessment instruments can only be interpreted in terms of identifying the strength and the weakness of the students or how much students know in the few content areas assessed. The instruments, thus, lack some degree of content validity.

An assessment task which lacks content validity is likely not to reflect the important content, skills and learning outcomes specified in the school's or district's curriculum framework and content standards (Nitko, 2001). This is reflected in the test papers where emphasis was placed on fewer content areas. This is because the test reflected the learning outcome of those few content areas taught.

Again, some of the questions in the paper A (multiple choice questions) were measuring trivial knowledge. These questions demanded the lowest form of thinking such that the test-wise student who does not have any knowledge of the material can answer these questions correctly. Nitko (2001) argues that most worthwhile learning involves students' using a combination of skills and content rather than using isolated skills or bits of content. This suggests that the assessment instruments do not measure worthwhile learning to some extent. It is evident that the assessment instruments lacked, to some extent, content representativeness and relevance.

The second part of the papers (Section B) required examinees to answer two questions out of four questions provided. It must be indicated that these four questions have different difficulty level and require a different level of cognitive operation to be able to attempt answering them. While some of the questions in this part were measuring knowledge, others were measuring comprehension. Examinees may end up answering different questions. The implication is that as Joan decides to attempt the first two questions, Isaac would be attempting the first and third questions. Francis can decide to answer first and fourth questions. Emmanuel would be tackling the second and third questions while Samuel might also answer the second and fourth question. Therefore, Francis' score cannot be compared to that of his peer who did not answer the same



questions he answered. This affects the soundness of the interpretation and use of the students' assessment result because the performance of examinees who have answered different questions with different difficulty level can never be compared.

### **3.2. Thinking Processes and Skills Represented**

An inspection of the test specification table for the multiple-choice items for all the papers indicate that the majority of the items measured lower-level skills. Thus, most of the items only required examinees to just recall facts. Few of the items measured comprehension and application. The essay part of the assessment instruments covered items on knowledge and comprehension. More specifically, all the sub-items on the compulsory questions were measuring knowledge. Surprisingly, it was found that the items in the essay part were measuring knowledge and comprehension with greater emphasis on knowledge.

The thorough account of the assessment tasks suggests that almost all the items were "recall" type of questions. This implies that examinees who engage in rote learning are those who will perform well and not necessarily those who have mastery over the material taught. Thus, the assessment instruments did not comprehensively assess different types of thinking skills. For an assessment result to be valid, the tasks should assess a student's ability to use strategies and processes that reflect how scholars in the discipline think (Nitko,2001). These assessment instruments deviated from Nitko's assertion. That is, only one lower-level cognitive process is greatly emphasized.

According to the Ministry of Education [MOE], Ghana (2012), the profile dimensions for an objective test for assessment should be 30% knowledge and 70% for comprehension and application. The assessment instruments did not meet the criteria given by MOE. The tests did not represent the kinds of thinking skills that the state's curriculum framework and performance standards suggest.

### **3.3. Reliability and Objectivity of the Test Items**

The assessment instruments had a longer test length (based on West Africa Examination Council's (WAEC) standard) which is likely to increase the reliability of the results. This is supported by Nitko (2001) who argued that "longer assessments (with more task per learning target) are more reliable than shorter assessments" (p. 41). The paper A of all the instruments had 40 multiple choice questions with a point for a question. This part of the test will be scored objectively which will improve the reliability of the result. The second part of the assessment instruments, where examinees had to select some number of questions out of a lot, is likely to be scored subjectively which might affect the reliability of the test results.

Some of the multiple choice questions had problems in its structure (i.e., syntax error, faulty stem, grammatical errors, and ineffective distractors) and this is likely to affect the reliability of the test results. These flaws are likely to provide clues for the students to get the right answer to the stem. The grammatical error, for instance, might also give the examinees a different understanding of the question. These problems are likely to affect the consistency of the test result because the response to these items would not reflect what the examinees know.

The options to the multiple-choice items in all the papers were arranged horizontally which is likely to affect the reliability of test scores. This is because there is the likelihood that examinees might waste a lot of time reading the options to the questions. This affects slow readers in their attempt to respond to the questions. However, the time allowed was sufficient for an average examinee to answer all the questions required. Again, the options to the multiple choice questions were not alphabetically arranged and this might lead to some identifiable patterns in the key to the questions.

The assessment instruments were not formatted very well. The items were clumsy with poor spacing. In some of the multiple choice questions, options to the same questions had inconsistencies. While some of the options to the same started with a capital letter some of them started with small letters. Again, the font theme and size were not consistent from the instruction to the last question.

### **3.4. Fairness to Different Students**

An evaluation of assessment tasks revealed that the tasks did not contain any form of information which gives a particular group of examinees advantage over others. This suggests that the assessment tasks were fair to the examinees with regards to gender, ethnic group, socio-economic background, among others. However, the clumsy nature of the assessment tasks might bring about unfairness to students who cannot read clearly when assessment tasks are clumsy. This was confirmed by Nitko (2001) that any assessment tasks must be fair to all examinees from all socio-economic background, ethnic group and language as well as students with disabilities who are mainstreamed in one class.

### **3.5. Practicality of the Assessment Task**

A critical evaluation of the assessment papers found that the time allocated was enough and allows the examinees to appropriately respond to the items. Even though time was allocated for the essay part of the test, scores for each item in the essay section were not indicated. This might affect the reliability of the assessment results. This is because the time spent on a particular question depends on the score allocated to it. This is supposed to be done to ensure that examinees do not waste much time on questions with low scores. This was explained by Amedahe and Asamoah-Gyimah (2016) that practicality is concerned with the necessary material and time allotted to the test. They claim that a tester should consider the following questions: Will students have enough time to complete the test and are there sufficient materials such as booklets or answer sheets, tables, chairs etc. available to present the test to complete the test effectively? The critical evaluation of the papers seems to suggest that sufficient answer booklet and time were made available for students to complete the test effectively.

## **4. CONCLUSION and RECOMMENDATIONS**

The evaluations of the tests obtained in the three core subjects revealed that teachers are weak in test construction. Even though some principles were done right, most of the critical issues which are related to validity and reliability were overlooked. This questions the validity of the results which would be awarded to these students. It is important for classroom teachers to be aware of the fact that the measurement of psychological constructs like academic achievement is a difficult thing to do. This is due to the complex and dynamic nature of human beings. However, there is the need for teachers to gather some information about students for decision making about curriculum, students and educational policy. This information is needed not only for teachers but also for parents, schools' management and policymakers. Because the information collected is used for decision making, it must be as accurate as possible. If a test with low validity and reliability are mostly used, then, inappropriate decisions are likely to be made.

The accuracy of classroom assessment results is very important but difficult to achieve. The complex nature of examinees, examination conditions, problems with test instruments and other factors reduces the validity of classroom assessment results. However, through careful planning of the test as well as adherence to principles in test construction, test assembling, test administration, scoring and result interpretation can help teachers to gather valid and reliable information about students. It, however, appears that some teachers do not have much knowledge in testing practices or do not simply adhere to the principles in testing. Although



Ghana, as a country, does not have a statewide standard in testing, it is important for the Ghana Education Service (GES) to train teachers in assessment (especially, testing practices). Thus, teachers are advised to adhere to the testing practices. It is highly recommended that head teachers take up the challenge of inviting resource persons from recognised academic institutions to organise workshops for teachers on a regular basis to sharpen their skills on effective test construction practices.

The authors of this paper, however, acknowledge that validity and reliability do not entirely rely on the instrument examined. Issues that have to do with the examination conditions such as invigilation, cheating, room ventilation, room lightning, among others also contribute to the variance in test scores. The authors did not adequately probe into some of these issues. We, therefore, recommend that further studies can go further to investigate some of these issues. It is vital, however, to say that the teacher plays a significant role in ensuring proper examination conditions. Again, caution should be taken not to generalise the findings of this study to a wider population.

### ORCID

Frank Quansah  <https://orcid.org/0000-0002-4580-0939>

### 5. REFERENCES

- Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Illinois: Waveland Press.
- Amedahe, F. K. (1989). *Testing practices in secondary schools in the Central Region of Ghana*. Unpublished Master's thesis, University of Cape Coast, Cape Coast.
- Amedahe, F. K., & Asamoah-Gyimah, E. (2016). *Introduction to measurement and evaluation*. Cape Coast: University of Cape Coast Press.
- Corbin, J., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3<sup>rd</sup> ed.). Thousand Oaks, CA: Sage.
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative and mixed methods approaches* (4<sup>th</sup> ed.). London: SAGE Publication Inc.
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Ohio: Cengage Learning Press.
- Curriculum Research and Development Division (CRDD) (1999). Investigation into student assessment procedures in public Junior Secondary Schools in Ghana. Accra: Ghana Education Service.
- Ebinye, P. O. (2001). Problems of testing under the continuous assessment programme. *J. Qual. Educ.*, 4(1), 12-19.
- Etsey, Y. K. A. (2003). Pre-service teacher's knowledge of continuous assessment techniques in Ghana. *Journal of Educational Development and Practice*, 1(1), 1-18.
- Etsey, Y. K. A. (2012). *Assessment in education*. Cape Coast: University of Cape Coast Press.
- Hamafyelto, R. S., Hamman-Tukur, A., & Hamafyelto, S. S. (2015). Assessing teacher competence in test construction and content validity of teacher made examination questions in commerce in Borno State, Nigeria. *Journal of Education*, 5(5), 123-128.
- Kazuko, J. W. (2010). *Japanese high school mathematics teachers' competence in real world problem solving*. Keto Academy of New York and Teachers College Columbia University.
- Ministry of Education [MOE] (2012). *Reports for 2012*. Accra: MOE Press.
- Nitko, J. A. (2001). *Educational assessment of students*. New Jersey: Prentice Hall.

- Ololube, N. P. (2008). Evaluation competencies of professional and non-professional teachers in Nigeria. *Studies in Educational Evaluation*, 34(1), 44-51.
- Onyechere, I. (2000). *New face of examination malpractice among Nigerian youths*. The Guardian Newspaper July 16.
- Quansah, F., & Amoako, I. (2018). Attitude of Senior High School (SHS) teachers towards test construction: Developing and validating a standardised instrument. *Research on Humanities and Social Sciences*, 8(1), 25-30.
- Tamakloe, E. K., & Amedahe, F. K. (1996). *Principles and methods of teaching*. Cantoment: Black Mask.

## An Effective Way to Provide Item Validity: Examining Student Response Processes

Omer Kutlu <sup>1</sup>, Hatice Cigdem Yavuz <sup>1,\*</sup>

<sup>1</sup> Department of Educational Measurement and Evaluation, Ankara University, Turkey

### ARTICLE HISTORY

Received: 01 August 2018

Revised: 11 December 2018

Accepted: 19 December 2018

### KEYWORDS

Cognitive interviews,  
Item validity,  
TIMSS,  
Response processes

**Abstract:** Studies based on response processes of individuals can provide information that supports the assessment and increases the validity of the items in the scale or tests. The purpose of this study is to present the extent to which the student response processes are effective in identifying and developing the characteristics of the items in an achievement test and in collecting validity evidence. For this purpose, 28 Turkish fourth-grade students were chosen, half were high-achieving students and the remaining half were low-achieving students. The items for the study were chosen from the Trends in International Mathematics and Science Study TIMSS 2007 and 2011 by taking into consideration several item characteristics. Before cognitive interviews, an interview guide was also prepared. In the study, it was determined that cognitive interviews, especially those conducted with the high-achieving students, can serve to develop item validity. In the cognitive interviews with the low-achieving students, information was gathered concerning how students who did not have specific knowledge measured with an item were able to respond to that item.

## 1. INTRODUCTION

One of the most important characteristics sought in tests used in education and psychology is validity. The ways of increasing validity by obtaining evidence of this characteristic are among the important issues that concern psychometricians. Validity is defined as “the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests” (American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), 1999, p. 9). For this reason, according to the Standards in Education and Psychology, validity in this context does not refer to the actual test but to the validity of the interpretations and evaluations made in consideration

---

CONTACT: Hatice Cigdem Yavuz ✉ [hcyavuz@ankara.edu.tr](mailto:hcyavuz@ankara.edu.tr) 📧 Department of Educational Measurement and Evaluation, Ankara University, Turkey

ISSN-e: 2148-7456 / © IJATE 2019

of the intended uses of the test. In this context, psychometricians search for evidence from different sources in relation to the validity of test scores (McDonald, 1999).

There are different opinions on the definition and classification of the term validity (see Sireci, 2007). Some researchers conceptualise validity within a general framework (AERA, APA, & NCME, 2014; Kane, 2013; Sireci & Foulkner-Bond, 2014; Sireci, 2007) and some suggest that validity cannot be interpreted generally (e.g. Borsboom, Mellenbergh, & van Heerden, 2004; Lissitz, & Samuelsen, 2007). Since tests used in psychology and education are developed depending upon specific purposes, people or conditions, it is not possible to develop a perfect test, which would serve all the required characteristics (Cronbach, 1984). From this point of view, in 1999 and 2014, the Standards in Education and Psychology (AERA, APA, & NCME, 1999, 2014) approached validity as a whole in the form of the types of validity without separation according to the types, such as content validity, criterion-related validity, and structural validity. The current study is based on this approach.

According to the Standards in Education and Psychology (AERA, APA, & NCME, 2014)), the ways of collecting the evidence of validity are divided into five sources of validity evidence; test content, response processes, internal structure, relations to other variables, and testing consequences. With these sources of validity evidence, the ways to obtain validity based on response processes has become an attention-grabbing subject in the field literature (Desimone & Le Floch, 2004; Padilla & Benítez, 2014; Ryan, Gannon-Slater & Culbertson, 2012). Evidence based on response processes is defined as "concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (AERA, APA, & NCME, 2014, p.12). In this way, thorough information is gathered regarding the cognitive processes shown by the respondents and response processes, and it is possible to determine to what extent these processes are in accord with the purposes of the test (Padilla & Benítez, 2014). In addition, together with the response processes it is possible to reveal how items are interpreted by individuals (DeWalt et al., 2007). Thus, studies based on response processes can provide information that supports the evaluation and increases the validity of the items in the scales or tests.

There are wide-ranging ways to obtain evidence based on the response processes, such as think aloud, focus groups and interviews (Padilla & Benítez, 2014). Among these, cognitive interview is a method composed of thinking aloud and verbal probing techniques (Willis, 2005). The role of the cognitive interview is important and useful in understanding the response processes of individuals (DeSimone & Le Floch, 2004; Ryan, Gannon-Slater, & Culbertson, 2012). With cognitive interviews, it is easier to discover which strategies individuals use and what they really think when responding to an item (Hopfenbeck & Maul, 2011). According to Desimone & Le Floch (2004), cognitive interviews can reveal mistakes in the item, different interpretations regarding the item, and the effect of the social desirability on the response to the item. Thus, measures to increase the validity of the items can be undertaken with the obtained data. Through cognitive interviews, it can be determined whether the items in measurement tools need to be reorganised (Conrad & Blair, 2004). In this context, cognitive interviews are used in pilot applications of scale development research and produce effective results (e. g. Johnstone, Figueroa, Attali, Stone, & Laitusis, 2013; Peterson, Peterson, & Powell, 2017; Snow & Katz, 2009; Wildy & Clarke, 2009).

In the literature, cognitive interviews are being investigated in various fields, such as health, education and social sciences. In their study, DeWalt et al. (2007) researched whether items in a scale related to a psychological structure were clear and understandable, and how individuals interpreted the items through the cognitive interview. With a similar purpose, Nicolaidis, Chienello and Gerrity (2011) showed through cognitive interviews that a 10-item scale is clear and understandable in terms of the focus group. In another study (Ding, Reay, Lee, & Bao,

2009), cognitive interviews were used in order to identify validity problems that could not be identified by experts and to present students' different perspectives towards the items. Ercikan, Arim and Law (2010) used the response processes of the students in order to examine the differential item functioning (DIF), which results from linguistic differences. With a similar purpose, Benitez and Padilla (2013) investigated DIF in student questionnaires used in the Program for International Student Assessment (PISA) and sought to reveal the possible sources of DIF by carrying out cognitive interviews with students. According to the findings of the study, words in some items were variously interpreted by students in different groups. Wildy and Clarke (2009) conducted cognitive interviews to test the preliminary test of the scale they used in their work and to test whether the meaning the scale writers attributed to an item was understood in the same way by the respondents. With a similar purpose, Ryan et al. (2012), identified measurement errors that had not emerged in other analysis methods using cognitive interviews to assess the validity of a scale. Ouimet, Bunnage, Carini, Kuh and Kennedy (2004) re-analysed the "College Student Report" tool developed for university students, within the framework of cognitive interview, focus group interview and expert opinions. The findings of the study show that these descriptive methods are important in improving the clarity, validity, appearance and reliability of the tool, as well as in revealing the strengths and weaknesses of the item.

Apart from the studies carried out on scales and questionnaires, cognitive interviews are also used on tests in the field literature. For example, Johnstone et al. (2013) used cognitive interviews to determine how students with disabilities interpreted the items in large-scale tests and to receive their feedback regarding the test items. In this framework, the differences were revealed between the responses of the disabled and non-disabled students in relation to the test items. In another study, cognitive interviews were carried out on the iSkills™ test developed by the Educational Testing Service (ETS) to measure the digital literacy skills of students (Snow & Katz, 2009). According to the findings of the study, evidence of validity was obtained for iSkills™ with regard to determining students' digital literacy. In their study, Noble, Rosebery, Suarez, Warren and O'Connor (2014) analysed the response processes of English Language Learner (ELL) students and non-ELL students through items gathered from a high-stakes science test. The findings of the study show that even though the ELL students have knowledge of the item, the linguistic features of the items led them to the wrong answer.

In the field of literature, it is seen that cognitive interviews are used on test items quite narrowly, and they are generally applied to improve the validity of the questionnaire or scales or to obtain evidence of validity. Thus, this study attempts to fill this gap in the field literature, and to show that cognitive interviews can also be applied to younger age groups and this application can be informative in terms of validity. Within this framework, the purpose of this study is to present the extent to which the student response processes are effective in identifying and developing the characteristics of the items in an achievement test and in collecting validity evidence. The research questions developed for this purpose are: What are students' response processes concerning (i) the necessity for the figure or table in the item root, (ii) the clarity of the text or figure given in the item root, (iii) the level of difficulty of the item, (iv) the level of knowledge given in the item, and (v) the reason for selecting the relevant choice in the item?

## **2. METHOD**

### **2.1. Participants**

The participants of this study were 24 Turkish fourth-grade elementary school students (10 girls, 14 boys) aged between 9 and 10. Cognitive interview studies are conducted with typically small sample sizes (Willis, 2015) thus the sample size in this study was enough to obtain well detailed student response processes. Moreover, the ability of items to measure the desired

feature without involving the group characteristics is very crucial. While gathering information about whether the items possess this feature, taking opinions from individuals with different characteristics will lead to more realistic and accurate information being obtained. For this reason, the schools to be involved in the research were carefully selected. Purposive sampling was employed in this study with the selection of two schools. After receiving permission from the authorities, one class was randomly selected from each school. In addition, participants were informed that participation was voluntary.

In this study, students were divided into two groups of high achieving and low achieving. One reason for this division is that students in these two different achievement levels would have different perspectives towards the items, which would frame their responses to the questions directed to them. The level of achievement of the students that the classroom teachers verbally indicated was taken into account. To determine the students' achievement level, the teachers were asked to consider students' academic achievement performance, participation in class, performing assignments and undertaking homework adequately, the level of interest and curiosity they have during the lesson, and briefly the students' performance in the classroom.

## 2.2. Data Collection Tools

The items for the study were chosen from the TIMSS 2007 and 2011 (TIMSS 2007 Assessment, 2009; TIMSS 2011 Assessment, 2013) and are presented in Appendix 1. In this study, students answered the Turkish version of these items which were translated and adapt in the framework of TIMSS assessments. Item statistics, cognitive domains and the scope of the items were taken into account in the selection of the items and attention was paid to the items being heterogeneous in terms of the related features. The TIMSS items were chosen on the basis of the item parameters being estimated according to item response theory and that the information regarding the items can be easily accessed. [Table 1](#) gives the characteristics regarding the items and the number of participants that gave correct answers.

**Table 1.** The characteristics regarding the items

| Item | Item type | Item discrimination | Item difficulty | Guessing parameter | Context   | Cognitive Domain | The number of participants that gave correct answers |               |
|------|-----------|---------------------|-----------------|--------------------|-----------|------------------|--|---------------|
|      |           |                     |                 |                    |           |                  | High-achieving                                       | Low-achieving |
| 1    | MC*       | .76                 | -1.64           | .22                | Biology   | Knowing          | 12   | 10            |
| 2    | MC*       | 1.12                | -1.14           | .26                | Biology   | Applying         | 11   | 7             |
| 3    | MC*       | .71                 | .14             | .22                | Chemistry | Knowing          | 9  | 8             |
| 4    | MC*       | .84                 | .39             | .18                | Chemistry | Reasoning        | 12   | 6             |
| 5    | OE**      | .53                 | 1.07            | -                  | Physics   | Reasoning        | 12   | 3             |
| 6    | MC*       | .75                 | -1.63           | .22                | Biology   | Applying         | 11   | 7             |
| 7    | OE**      | .95                 | 1.13            | -                  | Chemistry | Applying         | 8  | 3             |
| 8    | OE**      | 1.00                | .28             | -                  | Biology   | Knowing          | 6  | 1             |

\* MC: Multiple choice, \*\* OE: Open ended

[Table 1](#) shows that a total of six multiple-choice and three open-ended items were selected for use in the study. This distribution was chosen for the students to be able to respond to items during class time.

In this study, an interview form was another data collection tool. The interview form was composed of questions about the length, language, and the use of visual materials of the items. Techniques specific to the method of cognitive interviewing (Willis, 2015; Bowen, Bowen & Woolley, 2004) were employed in the preparation of the items in the interview form. After the interview form was prepared, the opinions of two experts in the field of educational measurement were taken. The experts were asked to express their views on the language and



expression characteristics of the questions, suitability for the age level, and the appropriateness of the extent of the questions. For pretesting, the interview form was applied to five students of the same age range from a class but were not participating in the study. The interview form was modified within the framework of the findings obtained from pretesting and the expert opinions.

### **2.3. Procedure**

In this study, first, the items were applied to the students, and then cognitive interviews were administered. Prior to the applications, information was given to the classroom teachers and students. It was explained to the students that their participation in the study would be kept confidential and their performance in response to the science items would not be shared with third parties apart from the researchers. It was also stated that participating in the test and interviews was voluntary.

After the required explanations in class were given by at least one researcher, the selected items were applied to the students. It took students approximately 30-40 minutes to respond to the items. After the test was applied, to avoid creating a negative impression towards the selection of the students, it was specified that interviews were to be conducted with the randomly selected students, and they were randomly summoned one by one from their classes from a list. During interviews, same protocols were followed with each group, and students were not told about the correctness of their choices on items as part of the interview.

Cognitive interviews were conducted on one-on-one basis with students in a different classroom in the school. Cognitive interviews are a combination of many techniques and provide a comprehensive understanding of how individuals comprehend and respond to items (Tourangeau, Rips & Rasinski, 2000). In this method, it is sought to determine the processes and thoughts that individuals experienced while replying to the items (Willis, 2015). Moreover, as pointed out by Willis, inferences are made regarding why the respondents in the cognitive interviews responded to a related item and how they responded in that way. Bowen et al. (2006) state that cognitive interviews can be carried out in the following way: First, the respondent is asked to read the item, and afterwards it is determined what meaning the item has for him and what he is being asked in the item. Why the student chose that answer is asked according to the student's response. Probing questions to be addressed in cognitive interviews can also reveal what the student actually thinks about an item when responding to that item. In this study, these steps were followed in each interview.

### **2.4. Data Analysis**

Before the analysis all interviews were transcribed, verbatim, yielding a total of 51 pages of written transcriptions. The data obtained from the interviews was analysed using descriptive analysis (Strauss and Corbin, 1990). In the context of this approach, first, the views of the students were placed in the predetermined themes and response categories. Determination of themes and response categories were established according to the answers of questions in the interview form. The placing the views into categories were conducted by the authors. This process was undertaken separately for each item. In the analysis of the data, the quotations that best explain the answers of the low and high achieving students were collated in accordance with each theme. In the study, the data belonging to four students selected impartially from the 24 students were also re-coded by an expert in educational measurement. With her re-coded data, the rater-reliability was determined as 96%.

## **3. FINDINGS**

### **3.1. Responses to the necessity of a Figure/Table**

The responses of students on the necessity of a figure/table to answer the items are given in order in [Table 2](#).

**Table 2.** The responses of the students on the necessity of a figure/table

| Item       | 2   |     | 4   |     | 5   |     | 6   |     | 7   |     |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|            | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A |
| Needed     | 6   | 10  | 10  | 8   | 7   | 6   | 6   | 4   | 11  | 7   |
| Not needed | 6   | 2   | 2   | 4   | 5   | 6   | 6   | 8   | 1   | 5   |

H-A: High achieving, L-A: Low achieving

According to [Table 2](#), most of the students stated that there was no need for a figure or table to answer item 6; yet, it was needed for item 7. The students who thought that a figure was needed to answer item 2 expressed this as, "I could see the characteristics of the walrus by looking at the figure". An examination of the items shows that being able to respond to items 4 and 7 depends on using the given figure. However, [Table 2](#) reveals that there were also students who stated that a figure/table was not necessary to respond to these items. The reason why a group of students did not consider a table necessary for responding item 4 is that the experiment column is not given in the table. In relation to this topic, one student commented, "I was confused by the table because the experiments were not stated one by one." According to the students' responses, the absence of a separate column that showed four experiments was considered to be "puzzling".

The students found the figure given in item 5 necessary since it "makes it easier to answer the item". Students presented their opinions as, "You understand their distance better in the visual", and "In the figure, A pulls from a farther distance". Generally, the figure in item 6 was considered as complicated by both high- and low-achieving students with comments such as "It is not clear that it is a frog", "the figure is confusing", and "the picture could be less blurred". For item 7, the higher-achieving students expressed their thoughts as, "the figure is necessary, I would not be able to understand if it were not for the figure" and "the figure is necessary, I did not understand the question when they were given one after the other", and the low-achieving students stated that "Items were long for me" and "Items were confusing".

### 3.2. Responses on text/figure clarity

[Table 3](#) presents the responses of students on the comprehensibility of the text/ figure in the items.

**Table 3.** The responses of the students on the comprehensibility of the text/ figure

| Item      | 1   |     | 2   |     | 3   |     | 4   |     | 5   |     | 6   |     | 7   |     | 8   |     |
|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|           | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A |
| Clear     | 12  | 7   | 9   | 7   | 12  | 8   | 9   | 6   | 11  | 6   | 11  | 7   | 8   | 4   | 7   | 4   |
| Not clear | -   | 5   | 3   | 5   | -   | 4   | 3   | 6   | 1   | 6   | 1   | 5   | 4   | 8   | 5   | 8   |

H-A: High achieving, L-A: Low achieving

According to [Table 3](#), the text in items 1, 3, 5 and 6 was considered as comprehensible by the high-achieving students. The choices in the item were the reason why items 1 and 6 were not understood by low-achieving students. For item 2, students stated that they did not understand the word "walrus". The responses of low achieving students to item 3 were: "The item is not completely expressed well" and "I could not understand the text very much, I did not understand the top part". For item 5, generally the comments of the low-achieving students regarding the text in the item were: "I did not understand the first sentence", "Puzzling", and "Text is too long". The low-achieving students stated the reason why they could not understand item 7 was as follows: "I could not understand the cups", "x, y cup names", and "I could not understand the text very much". For the same item, the high-achieving students gave these responses: "I



was confused by the expression" and "I did not understand how he/she placed the container without taking out the glass in the figure".

The text of item 8 was not found to be understandable by a large number of both the high- and low-achieving students. The low-achieving students expressed their views as, "I perceived it as we can see plants and seed from a distance", "I did not understand the word 'from far away'", and "The text is complicated, providing a figure would be better". The views of the high-achieving students were given as "puzzling", "I understood it when I read it twice", and "The words 'far away' made me confused".

Generally, as can be understood from these responses, when making decisions about data to determine whether items are clear and understandable, the group which responded correctly to the item most of the time should be taken into consideration. Taking the students' responses and Table 3 into account, for the less successful students, the comprehensibility of the text becomes more difficult as the length of the text increases.

### 3.3. Students' responses on item difficulty

Table 4 presents the responses of students about difficulty of the item.

**Table 4.** The responses of the students about difficulty of the item

| Item           | 1   |     | 2   |     | 3   |     | 4   |     | 5   |     | 6   |     | 7   |     | 8   |     |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|                | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A |
| Easy           | 8   | 8   | 12  | 8   | 8   | 5   | 6   | 6   | 10  | 8   | 11  | 8   | 5   | 9   | 5   | 6   |
| Medium         | 4   | 3   | -   | 2   | 4   | 4   | 3   | 3   | -   | 2   | -   | 1   | 5   | -   | 3   | 5   |
| Most difficult | -   | 1   | -   | 2   | -   | 3   | 3   | 3   | 2   | 2   | 1   | 3   | 2   | 3   | 4   | 1   |

H-A: High achieving, L-A: Low achieving

According to Table 4, the high-achieving students found item 2 the easiest whereas items 7 and 8 were found to be the most difficult. For the low-achieving students, item 7 was the easiest and item 3 was the most difficult. The high-achieving students explained why they generally found item 1 easy or of medium difficulty as, "This information is in cartoons", "I learned this in the documentary", "Because I've read in books", "Among the choices, it made the most sense" and "I could not decide between B and C" whereas the low-achieving students stated, "I had a bird and therefore I know it", "I am good with animals", and "I know about birds". The high-achieving students gave the reasons for finding item 2 easy as, "I learned it from a documentary" and "I learned from documentaries and books". Those among the low-achieving students who found item 2 easy explained, "I learned it from the TV" and "Choices are nonsense", and the remaining students expressed their view as, "I have never heard of a walrus", "I do not know what palette is", and "Walrus is a different animal".

For item 3, the higher-achieving students who found the item easy expressed their view as, "We saw it in class" and "It melts instantly because it is hot", and those who found it to be of medium difficulty stated, "I was confused because first it was cold then warm and hot." Those of the low-achieving students who found this item easy commented, "I knew it, it dissolves in cold when the heat is high", and those who found it to be of medium difficulty or difficult stated, "warm and hot water puzzled me". Those who found item 5 easy generally expressed their view as, "I solved it with the help of other options" and "the other options are meaningless". Those who found the item to be of medium difficulty or difficult stated, "it is because of the table" and "the table confused me".

Both high- and low-achieving students generally found item 5 easy. The high-achieving students gave reasons, such as "the answer is written directly" and "the answer is definitely

clear", and the low-achieving students gave responses; for example, "the figure made it easier" and "the picture made it easier".

Item 6 was also generally found to be easy by the high- and low-achieving students. The higher-achieving students responded as, "there were things that could not be possible in other options" and "I saw it in the movies that the frogs leave" whereas the responses of the low-achieving students were "I love animals", "Options are easy", and "the answer is apparent".

Both high- and low-achieving students generally found item 7 difficult or of medium difficulty giving responses, such as "I did not understand the complicated text much, it was a little long", "I do not understand how they put the x in the box", "I was confused because of the figures", and "I had difficulty because of the text". The high- and low-achieving students who found this item easy commented, "Figures made it easier", "I learned it in science lessons", and "the options made it easier".

Item 8 was also found to be difficult or moderately difficult by the majority of the high and low achieving students. The students expressed their views as: "The sentence is complicated, it is not clear", "it would be better if they added a figure", "I did not understand the seed in the text", "I did not understand the words 'far away' in the text". The students who considered the item to be easy gave responses such as: "I imagined, estimated in my mind", "I like the plant kingdom", "I see it in villages", "I see it in documentaries" and "I pictured it in the soil".

### 3.4. Responses regarding the level of knowledge given in the item

In Table 5, the responses on the level of knowledge given in the item are given according to responses of the high- and low-achieving students.

**Table 5.** The responses of the students on the level of knowledge given in the item

| Item          | 1   |     | 2   |     | 3   |     | 4   |     | 5   |     | 6   |     | 7   |     | 8   |     |
|---------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Student group | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A |
| Enough        | 12  | 9   | 11  | 9   | 11  | 7   | 12  | 8   | 11  | 8   | 11  | 8   | 10  | 8   | 6   | 3   |
| Not enough    | -   | 3   | 1   | 3   | 1   | 5   | -   | 4   | 1   | 4   | 1   | 4   | 2   | 4   | 6   | 9   |

H-A: High achieving, L-A: Low achieving

According to Table 5, the high-achieving students indicated that items 1 to 7 contained sufficient information to find the answer. The low-achieving students generally expressed their views as bird features could be given in item 1, features about walrus could be given in item 2, figures could be given in item 3 and 7, a table was missing in item 4, and the item root was inadequate in item 6. To solve item 8, both the high- and low-achieving students stated that there was not enough information. The opinions of the high-achieving students were: "I think there is a need for some sentences and some figures", "a little information can be added", "there is no information, direction", "it could have been expressed in a different way", and "a figure can be added", and the low-achieving students shared these views with their comments of "how far away?", "I was puzzled by the words 'far away'", "there could be a figure", "'far away' is not something that can be expressed", and "it is a little hard; it could have been simplified."

### 3.4. Responses on the students' selection of the relevant item option

The responses of the high- and low-achieving students on the reasons for selecting the relevant option are given in Table 6. According to this table, the reason for students' selection of an option they considered to be the correct answer was generally related to the other options. Thus, the students found the correct answer by eliminating the other options that were given before questioning the information in the item. For item 3, most of the high-achieving students found

the right answer using their knowledge. In addition, when finding the right answer, both the high- and low- achieving students were assisted by the other choices with most assistance provided by item 6.

**Table 6.** The responses of the students on the reasons for selecting the relevant option

| Item          | 1   |     | 2   |     | 3   |     | 4   |     | 6   |     |
|---------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|               | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A | H-A | L-A |
| Knowledge     | 6   | 5   | 3   | 3   | 10  | 6   | 8   | 1   | 2   | 1   |
| Other options | 6   | 7   | 9   | 7   | 2   | 6   | 4   | 8   | 9   | 7   |

H-A: High achieving, L-A: Low achieving

Generally, the high-achieving students who selected the correct answer by benefiting from the options in all items expressed their opinions as, “it was clear to me because of the other choices”, “I eliminated the other options”, “the other options cannot be correct because they are unreasonable”, “A was the most meaningful option”, and “it was quite clear in this option” while the high-achieving students who selected the correct answer by based on their knowledge commented, “I learned it from books and documentaries,” “it does not melt in cold water; I know it from tea”, and “because the temperature and the mixture are always the same”. The opinions of the low-achieving students were similar to these views. Among these students, those who marked the item based on the knowledge they possessed explained, “the layer of fat will protect the animal” and “heat melts sugar” while those who first eliminated the other options gave opinions, such as “the options give a lot of clues”, “by eliminating the options”, and “it makes more sense than the other options”.

#### **4. DISCUSSION and CONCLUSION**

The aim of this study was to show that the responses of the fourth-grade students to some items chosen from the TIMSS application could be used as validity evidence and to increase item validity. In the study, opinions were obtained from the high- and low-achieving students, and it was determined that cognitive interviews, especially those conducted with the high-achieving students, can serve to develop item validity. In the cognitive interviews with the low-achieving students, information was gathered concerning how students who did not have specific knowledge measured with an item were able to respond to that item.

Generally, this study shows that students, like experts, can have a role in providing evidence for item validity and increase validity. The findings of this study were found to be similar to those of other studies in the field literature (Ercikan et al., 2010; Nicolaidis, Chienello & Gerrity, 2011; Noble et. al, 2014). Validity evidence obtained through this research show that students are important as consulting experts in the processes of test development and adaptation; thus, this is an effective method to find solutions to validity issues. Even though cognitive interviews are time-consuming and costly (Desimone & Le Floch, 2004), it appears that this type of studies would be useful in improving the validity of the items and the tests. Although referring to experts’ opinions is an important and widespread method, it is clear that student’s response processes also provide validity evidence since experts cannot possibly have any knowledge about the student's response processes (Benitez & Padilla, 2013).

In situations where students’ high-level cognitive skills such as problem-solving and critical thinking are to be measured, verbal, numerical materials or materials; e.g. tables and figures are often used in the root of the developed item; therefore, attention should be paid to the selection of these materials (Beddow, Elliot & Kettler, 2013). According to the findings regarding the necessity of the materials in the items selected for this study, the students stated that some materials are necessary for some items and unnecessary for others. Considering that materials

that do not contribute to the solution of an item may negatively affect the item validity (Linn & Gronland, 1995, Nitko & Brookhart, 2007), it is clear that the visuals of such items need to be modified. For example, the material used in item 6 in the current study did not contribute to the student(s)' response to that item. In this sense, it is suggested that visuals that are not essential for students to respond to an item should not be included as item materials.

There may be a different approach to the inclusion of material in items which aim to measure psychological structures to ensure that such items are developed without bias (Haladyna, 1997; Osterlind, 2002). Since international measurement applications such as TIMSS and PISA are applied in numerous different cultures, material can sometimes be used in the root of the item even if it is not necessary, and the opinions that students expressed for item 2 in the current study support this. Some students stated that they benefited from the picture since they had not seen a walrus before, and a picture was needed in order to solve the item. In this sense, considering the cultural and socioeconomic characteristics of the groups to which the items will be applied is also significant regarding the validity of the item. The interviews with the students revealed that the material in item 5 provided a clue for the low-achieving students to find the correct answer. However, the materials in the root of the item should not give the students a clue to the correct answer (Haladyna, 1997). It is important to recognise that reliability and validity in tests depend on the selection of the items (Linn, 1989); therefore, as an example, the material in item 5 should be changed or removed.

In tests that measure psychological characteristics, the comprehensibility of the items is important in terms of language and expression. In the processes of writing items, the language used in the items and generally in the whole test must be clearly written and comprehensible, obeying spelling and punctuation rules (Haladyna, 1996; Osterlind, 2002). According to the student opinions, the items selected for this study from TIMSS appeared to have problems in their Turkish language. In addition, in order for the material in the base of the item to be understandable, the class and age level to which the item will be applied should also be considered (Linn & Gronland, 1995). The current study revealed that in the cognitive interviews, item 7 was more complicated than it should be for fourth-grade students. The x and y letters used in item 7 were confusing for the students. Furthermore, the word 'walrus' in item 2 also prevented the students from comprehending the item.

In tests, such as TIMSS, which have been translated from the language in which they are originally developed into a different language, the process of translation is important in the sense that gives the meaning which the item desires to measure, and represents the characteristics in the item. From this point of view, regarding the findings related to the items in the current study, it can be seen that in particular, the meaning of item 8 was lost. Regardless of whether the students were lower or higher achieving, it is necessary to create a situation in which all students understand the items and only the ones who have the knowledge and skills related to the item can answer it correctly (Linn & Gronland, 1995). The findings obtained from the current study show that there are problems in this respect; thus, it is suggested that changes are made to the items based on the opinions of the students.

Significant findings about the difficulty level of the items were obtained in the study. According to the cognitive interviews, the findings indicated that the items based on remembering a related specific knowledge were easy for the students to solve; items such as comprehension, problem solving, and critical thinking were more difficult because they measure information in a more complex way. In addition, the students attributed the ease of solving some of the items to the distractors not being related to the correct answer. Furthermore, extra-curricular resources such as documentaries and books also contributed to the response of the items.

It is observed in the current study that students generally had difficulty in open-ended items. Similarly, in a study conducted by Johnstone et al. (2013), the students achieved more correct

answers in multiple-choice tests. Another reason for the Turkish students having difficulties with open-ended items is that they are not familiar with this type of items in their instructional programme. To resolve this problem, the students' responses can be consulted to determine the level of difficulty of the items to be included in the tests. In large-scale international tests, such as TIMSS and PISA, differences in socioeconomic and sociocultural characteristics of the participating countries can also affect the difficulty levels of the items. The reason for the distractors of some of the items not being connected to the correct option may be due to the differences between countries and the results of such examinations being a particular concern to the educational policy makers. In such measurement applications, the percentage of items to which there is a correct response is a particular concern for educational policy makers since it leads to various inferences being made about the countries. Hopfenbeck and Maul (2011) emphasise the need to take care over comparisons of countries with the information gained from these measures.

When the findings in the current study are analysed in relation to the sufficiency of the information required for the solution of the items, the high-achieving students generally found the given information sufficient with the exception of item 8. It is understood from the opinions of the low-achieving students that the information given in the item was not adequate. Taking the other responses given to item 8 into account, it was problematic in many respects. Therefore, it would be wrong and biased to compare Turkish students with other students who responded to this item because this item does not operate in the same way for the Turkish-speaking students. Mistakes originating from translation can be observed in tests such as TIMSS and PISA (see Goldstein, 2008). The diversity of participating countries taking these tests and that the tests do not follow the various curricula in the schools makes the item writing process difficult. In this framework, the tests should be based on common learning topics, and preliminary research on the test content should be undertaken by obtaining opinions from the participating countries. This process will be significant in developing the scientific accuracy of the items and the general validity of the test.

Considering the students' reasons for selecting an option for the items addressed to them in the study, it is seen that writing multiple-choice questions is as important as writing the item base. In this context, various precautions should be taken in writing the options in multiple-choice items (Haladyna, 1997, Nitko & Brookhart, 2007). Moreover, the response process of a multiple-choice item depends on the characteristics of the student as much as the characteristics of the item itself. While some students read the item and look for the expression in the choices they think is the correct response, some students try to obtain the answer by comparing the choices with each other after reading the item (Pehlivan Tunç & Kutlu, 2014). In this sense, some students can develop test-wiseness behaviour when responding to the items. This situation decreases the validity of the test, and in order to avoid such strategies, measures should be taken in the development and review of the items (Townes, 2014). It was also determined in this study that distractors should be rational and consistent with the context especially in multiple-choice tests (Osterlind, 2002).

When the findings obtained from this study are considered as a whole, taking into account student responses in item writing and undertaking the corresponding improvement of the items will provide significant contributions to the validity of the tests. There are only a few studies on this subject; therefore, more studies being conducted will help gain different perspectives in test development processes. In future studies, it would be appropriate to use different item types, benefit from a larger number of items and refer to the opinions of students who are studying at different levels. In addition, researchers can conduct intercultural cognitive interview studies to support the development studies of scales and tests, such as TIMSS, which are applied in different cultures. This study has some limitations and premises. In this sense, the results of this



study need to be evaluated within this framework. A significant limitation is that the students were fourth graders, and they had not participated in this type of interviews before. For this reason, some data loss was experienced in some questions. Moreover, this study was conducted with only 24 students. Another limitation is that the study used a total of eight multiple-choice and three open-ended items. The assumptions that students responded to the questions without being affected by social desirability and that students responded to questions solely based on their own knowledge were accepted as the premises of the study.

### Acknowledgement

This study was presented at the 10th International Test Commission (ITC) Conference in 1 - 4 July 2016 in Vancouver, Canada.

### ORCID

Ömer KUTLU  <https://orcid.org/0000-0003-4364-5629>

Hatice Çiğdem YAVUZ  <https://orcid.org/0000-0003-2585-3686>

### 5. REFERENCES

- AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2013). Test accessibility: Item reviews and lessons learned from four state assessments. *Education Research International*, 2013, 1-12. doi:10.1155/2013/952704
- Benitez, I., & Padilla, J. L. (2013). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research*, 8(1) 52-6. doi: 10.1177/1558689813488245
- Borsboom, D., Mellenbergh, G. J., & Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061-71. doi: 10.1037/0033-295X.111.4.1061
- Bowen, N. K., Bowen, G. L., & Woolley, M. E. (2004). Constructing and validating assessment tools for school-based practitioners: The Elementary School Success Profile. In A. R. Roberts & K. Y. Yeager (Eds.) *Evidence-based practice manual: Research and outcome measures in health and human services* (pp. 509-517). New York: Oxford University Press.
- Conrad, F., & Blair, J. (2004). Aspects of data quality in cognitive interviews: The case of verbal reports. In S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, et al. (Eds.) *Questionnaire development, evaluation and testing methods* (pp. 67-88). New York: Wiley.
- Cronbach, L. J. (1984). *Essentials of psychological testing*. NY: Harper.
- Desimone, L., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis*, 26(1), 1-22. doi:10.3102/01623737026001001
- DeWalt, D. A., Rothrock, N., Yount, S., et al. (2007) Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care*, 45(1), 12-21. doi: 10.1097/01.mlr.0000254567.79743.e2
- Ding, L., Reay, N. W., Lee, A., & Bao, L. (2009). Are we asking the right questions? Validating clicker question sequences by student interviews. *American Journal of Physics*, 77(7), 643-650. doi:10.1119/1.3116093

- Ercikan, K., Arim, R., & Law, D. (2010). Application on think aloud protocols for examining and confirming sources of differential item functioning identified by experts review. *Educational Measurement: Issues and Practices*, 29, 24-35. doi:10.1111/j.1745-3992.2010.00173.x
- Goldstein, H. (2008). *How may we use international comparative studies to inform education policy*. Retrieved from <http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/how-useful-are-international-comparative-studies-in-education.pdf>
- Haladyna, T. M. (1996). *Developing and validating multiple-choice test items*. NJ: Lawrence Erlbaum associates, publishers.
- Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. USA: Allyn & Bacon.
- Hopfenbeck, T. N., & Maul, A. (2011) Examining evidence for the validity of PISA Learning Strategy Scales based on student response processes. *International Journal of Testing*, 11(2), 95-121. doi: 10.1080/15305058.2010.529977
- Johnstone, C., Figueroa, C., Yigal, A., Stone, E., & Laitusis, C. (2013). *Results of a cognitive interview study of immediate feedback and revision opportunities for students with disabilities in large scale assessments (Synthesis Report 92)*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. doi: 10.1111/jedm.12000
- Linn, R. L. (1989). *Educational measurement*. NJ: American Council on Education and Macmillan Publishing Company.
- Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). Englewood Cliffs, New Jersey; Prentice Hall.
- Lissitz, W. R., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437-448.
- Nicolaidis, C., Chienello, T., & Gerrity, M. (2011). The development and initial psychometric assessment of the centrality of Pain Scale. *Pain Medicine*, 12, 612-617.
- Nitko, A. J., & Brookhart, S. M. (2007). *Educational assessment of students* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Noble, T., Rosebery, A., Suarez, C., Warren, B., & O'Connor, M. C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education*, 27(4), 248-260.
- Osterlind, S. J. (2002). *Constructing test items: Multiple-choice, constructed-response, performance and other formats*. New York: Kluwer Academic Publishers.
- Ouimet, J. A., Bunnage, J. C., Carini, R. M., Kuh, G. D., & Kennedy, J. (2004). Using focus groups, expert advice, and cognitive interviews to establish the validity of a college student survey. *Research in Higher Education*, 45(3), 233-250.
- Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26, 136-144. doi: 10.7334/psicothema2013.259
- Pehlivan Tunç, E. B., & Kutlu, Ö. (2014). Investigation of Answering Behaviour in Turkish Test. *Journal of Measurement and Evaluation in Education and Psychology*, 5(1), 61-71.
- Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive interviewing for item development: Validity evidence based on content and response processes, *Measurement and Evaluation in Counseling and Development*, 50(4), 217-223, doi: 10.1080/07481756.2017.1339564

- 
- Ryan, K., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving survey methods with cognitive interviews in small- and medium-scale evaluations. *American Journal of Evaluation, 33*(3), 414-30. doi:10.1177/1098214012441499
- Sireci, S., & Faulkner-Bold, M. (2014). Validity evidence based on test content. *Psicothema, 26*, 1, 100-107. doi: 10.7334/psicothema2013.256
- Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481. doi: 10.3102/0013189X07311609
- Snow, E. & Katz, I. (2009). Using cognitive interviews to validate an interpretive argument for the ETS ISKILLS assessment. *Communications in Information Literacy, 3*(2), 99-127.
- Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage Publications, Inc.
- Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. NY: Cambridge University Press.
- Towns, M. H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education, 91*(9), 1426-1431. doi: 10.1021/ed500076x
- TIMSS 2007 Assessment. Copyright © 2009 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
- TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
- Wildy, H., & Clarke, S. (2009). Using cognitive interviews to pilot an international survey of principal preparation: A Western Australian perspective. *Educational Assessment, Evaluation and Accountability, 21*(2), 105-117. doi: 10.1007/s11092-009-9073-3
- Willis, G. (2015). *Analysis of the cognitive interview in questionnaire design (understanding qualitative research)*. NY: Oxford University Press.
- Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.



## Appendix 1 – Selected Items

### Item 1

Most birds sit on their eggs until they hatch. Which of these is the most important reason why birds sit on their eggs?

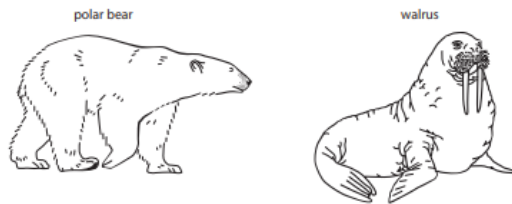
- a) to keep the eggs inside the nest
- b) to keep the eggs warm
- c) to protect the eggs from wind
- d) to protect the eggs from the rain

### Item 3

Sue measured how much sugar would dissolve in a cup of cold water, a cup of warm water, and a cup of hot water. What did she most likely observe?

- a) The cold water dissolved the most sugar.
- b) The warm water dissolved the most sugar.
- c) The hot water dissolved the most sugar.
- d) The cold water, warm water and hot water all dissolved the same sugar.

### Item 2



Polar bears and walrus look very different, but both can survive in extremely low temperature. Polar bears have a thick coat of fur that helps it keep itself warm. On the other hand, walrus have no fur.

What do walrus have to keep them warm?

- a) Fat layers
- b) Tusks
- c) Whiskers
- d) Flippers

### Item 4

Maria designed an experiment using salt and water. The results of her experiment are shown in the table.

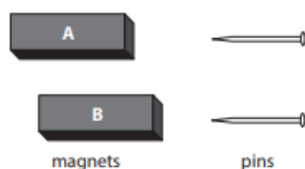
| Amount of Salt Dissolved | Water Volume | Water Temperature | Was Mixture Stirred? |
|--------------------------|--------------|-------------------|----------------------|
| 15 grams                 | 50 ml        | 25° C             | Yes                  |
| 30 grams                 | 100 ml       | 25° C             | Yes                  |
| 45 grams                 | 150 ml       | 25° C             | Yes                  |
| 60 grams                 | 200 ml       | 25° C             | Yes                  |

What was Maria studying in her experiment?

- a) How much salt will dissolve in different volumes of water.
- b) How much salt will dissolve at different temperatures.
- c) If stirring increases how fast salt will dissolve.
- d) If stirring decreases how fast salt will dissolve

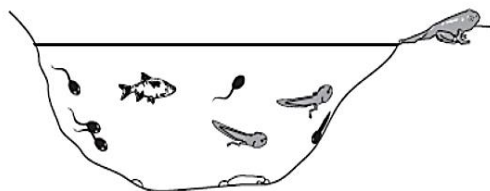
**Item 5**

Betty has two magnets (A and B) and two metal pins that are the same. She slides Magnet A along a table until a pin is attracted to the magnet. She slides Magnet B along a table until a pin is attracted to the magnet.



She finds that Magnet A attracts the pin from 15cm and Magnet B attracts the pin from 10cm. Steven says that both magnets are equally strong. Do you agree? Explain your answer.

**Item 6**

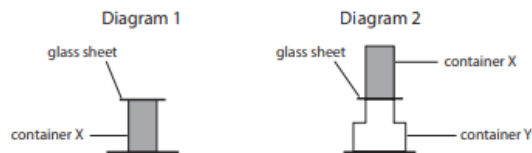


Melissa found some tadpoles and fish in a pond as shown above. How did the tadpoles get there?

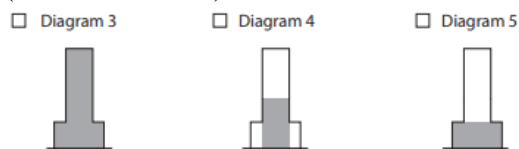
- They hatched from eggs laid by fish in the pond.
- They formed from mud at the bottom of the pond.
- They were made from materials dissolved in pond water.
- They developed from eggs laid by frogs in the pond

**Item 7**

Diagram 1 shows a container X that is filled with a material that could be a solid, liquid, or gas. The container has been sealed with a glass sheet. Container X is placed upside down on an empty container Y, as shown in Diagram 2.



The glass sheet is removed. A. Which of the diagrams below shows what you would see if the material in container X is a gas? (Check one box.)



B. Explain your answer.

**Item 8**

Seeds from a plant can end up a long way away from the plant. Describe one way that this can happen.

## Can Factor Scores be Used Instead of Total Score and Ability Estimation?

Abdullah Faruk Kilic  <sup>1,\*</sup>

<sup>1</sup> Department of Educational Measurement and Evaluation, Hacettepe University, Ankara, Turkey

### ARTICLE HISTORY

Received: 11 July 2018

Revised: 20 October 2018

Accepted: 07 December 2018

### KEYWORDS

Factor Score,  
Ability Estimation,  
Classical Test Theory,  
Item Response Theory,  
Total Score

**Abstract:** The purpose of this study is to investigate whether factor scores can be used instead of ability estimation and total score. For this purpose, the relationships among total score, ability estimation, and factor scores were investigated. In the research, Turkish subtest data from the Transition from Primary to Secondary Education (TEOG) exam applied in April 2014 were used. Total scores in this study were calculated from the total number of correct answers given by individuals to each item. Ability estimations were obtained from a three-parameter logistic model chosen from among item response theory (IRT) models. The Bartlett method was used for factor score estimation. Thus, the ability estimation, sum, and factor scores of each individual were obtained. When the relationship between these variables was investigated, it was observed that there was a high-level, positive, and statistically significant relationship. In the result section of this study, as variables have a high-level relationship, it was suggested that since variables could be used interchangeably, factor scores should be used. Although the total scores of individuals were equal, there were differences in terms of factor score and ability estimations. Therefore, it was suggested that item response theory assumptions were not met, or factor scores should be used when the sample size is small.

## 1. INTRODUCTION

Scales or achievement tests are commonly used to measure psychological traits of individuals. Based on the data obtained from such measurement tools, scores (ability scores) are obtained for individuals with different methods. There are various scoring methods for determining the level of individuals being measured. Classical test theory (CTT), item response theory (IRT), and factor scores are among these methods.

In classical test theory (CTT), the total number of correct answers given by individuals to items is generally preferred as a scoring method. Typically, the observed scores of individuals is referred to as the total number of correct answers to items (de Ayala, 2009; Price, 2017). Additionally, in CTT, options and items can be weighted, and different scoring methods can be

---

CONTACT: Abdullah Faruk Kılıç ✉ [afarukkilic@windowslive.com](mailto:afarukkilic@windowslive.com) 📧 Department of Educational Measurement and Evaluation, Hacettepe University, Ankara, Turkey

ISSN-e: 2148-7456 / © IJATE 2019

used. However, it is known that the contribution of these methods in terms of reliability and validity is not high, and efforts are higher than contributions (Gulliksen, 1950).

In IRT, unlike CTT, a non-linear relationship is formed between the answers of individuals to items and their abilities (DeMars, 2010; Hambleton & Swaminathan, 1985). In IRT, item discrimination and item difficulty can affect the estimation of abilities of individuals depending on the selected IRT model. In unidimensional IRT models, there are assumptions such as unidimensional, local independence, and the S-shape of item characteristic function (de Ayala, 2009). By investigating which IRT model (Rasch, one-, two-, and three-parameter logistic models) fits the data, abilities are estimated via that IRT model.

Another scoring method is factor score estimation. Factor score estimation can be divided into two main sections: 1) nonrefined methods and 2) refined methods. Among nonrefined methods, sum methods calculated based on CTT are included (DiStefano, Zhu, & Mîndriță, 2009). The easiest way to obtain factor scores is to sum raw scores of items relating to item loadings (Comrey & Lee, 1992). An important point to be considered here is to subtract item scores when factor loadings are negative (DiStefano et al., 2009). Another method that is classified as nonrefined and used for estimating factor scores is to sum certain items with factor loadings above a certain threshold. Another method is to use the sum of the standardized scores. Additionally, the sum can be calculated with weighted factor loadings of items. However, in this case, the measurement tool has to be unidimensional (DiStefano et al., 2009).

Refined methods applied in obtaining factor scores can be listed as the regression method, the Bartlett method, and the Anderson-Rubin method. In the regression method, the least squares method is used to obtain the factor score for each individual regarding factor or component. Factor scores are used as dependent variables in regression equations. In the Bartlett method, only common factors influence factor scores. In this method, squares of error variance of variables are minimized. It has been stated that the Bartlett method is unbiased for estimating real factor scores (Hershberger, 2005). The Anderson-Rubin method (Anderson & Rubin, 1956) is derived from the Bartlett method. In this method, factor scores are obtained as unrelated to both other factors and to each other. This method involves more complex calculation processes than the Bartlett method where the factor score is orthogonal, the average is 0, and the standard deviation is 1 (DiStefano et al., 2009).

Factor score estimation methods have some advantages. For example, since the correlation between factor score and factors is maximum in the regression method, it has been stated that more valid results are obtained. On the other hand, the Bartlett method can estimate factor score in an unbiased way. In the Anderson-Rubin method, factor scores obtained from two orthogonal factors can be unrelated (DiStefano et al., 2009). It can be said that factor scores obtained from these methods have a high-level relationship (Hershberger, 2005; Horn, 1965). There is also a factor score indeterminacy problem while estimating factor scores. When the total of common and unique variance exceeds the number of items, since the matrix formed from these elements is not a square matrix, the inverse of the matrix cannot be calculated. In this case, factor indeterminacy arises (Grice, 2001).

Generally, total scores are used in scales or achievement tests to decide about individuals. If analysis is conducted based on IRT, the ability parameter is estimated. Estimating factor scores is limited in studies. However, when the total score is considered, item characteristics have no effect on the ability of individuals. The fact that item characteristics influence individuals' ability estimation is seen as an advantage that IRT has over CTT (DeMars, 2010). Therefore, ability estimations are affected by item characteristics. Thus, the effects of items with strong psychometric properties on ability estimates are different. A similar situation is observed while estimating factor scores. Factor scores can be calculated by using factor loadings, unique

variances, regression weights, eigenvalues, and eigenvectors (DiStefano et al., 2009). Therefore, the abilities of individuals can be estimated more accurately.

When the literature is reviewed and item and ability parameters obtained from CTT and IRT are compared, there are studies analyzing parameter invariance (Akyıldız & Şahin, 2017; Bulut, 2018; Çakıcı-Eser, 2013; Cappelleri, Jason Lundy, & Hays, 2014; Çelen & Aybek, 2013; İlhan, 2016; Macdonald & Paunonen, 2002; Stage, 1998a, 1998b; Xu & Stone, 2012). In these studies, generally parameters obtained from IRT and CTT are compared and invariance property is generally obtained in IRT. But it can be said that parameter invariance is hold in CTT with larger samples. However, there are studies considering factor score estimation methods (DiStefano et al., 2009; Green, 1976; Hershberger, 2005; Horn, 1965; Williams, 1978). In these studies, factor score estimations are introduced and, in particular, factor score indeterminacy is emphasized. In addition to these studies, the relationship between factor score and scale scores (Fava & Velicer, 1992) or factor scores obtained from different factor extraction methods were compared with the scale score (Fava & Velicer, 1992; Grice, 2001; Velicer, 1976). These studies were generally conducted as a simulation study. In the current study, the aim is to investigate whether factor scores can be used instead of ability estimation and total score with high-stakes test data. Therefore, the current study investigated whether factor scores can be used instead of ability estimation and total score. Accordingly, in this study, the answer to the question “According to the relationship between total score, ability estimation, and factor score, can factor scores be used instead of ability estimation and total score?” was investigated. Therefore, with the help of the relationship between these variables, suitability of scores for deciding about individuals was discussed.

## **2. METHOD**

In this study, conducted to analyze the relationship between total score, ability estimation, and factor score, the research design was a relational study. In relational studies, relationships and connections are investigated (Büyüköztürk, Kılıç-Çakmak, Akgün, Karadeniz, & Demirel, 2013; Fraenkel, Wallen, & Huyn, 2012). In correlational studies among relational studies, correlations between two or more variables/scores are analyzed (Creswell, 2013). In this study, since the relationship between total score, ability estimation, and factor score was analyzed, the correlational research method was selected from among relational research methods.

### **2.1. Study Group**

In this study, data obtained from the Turkish Test in Transition from Primary to Secondary Education (TEOG) exam applied in April 2014 were used. Accordingly, from among 1,271,284 students, 10,000 students were sampled using simple random sampling. Based on this information, it can be stated that the sampling method of this study was simple random sampling (Büyüköztürk et al., 2013). Data cleaning was applied by analyzing a 10,000-sample data set. Accordingly, data of individuals with repetitive answers or who gave the same answers to each question were deleted and analyses were conducted on 9,773 student data.

### **2.2. Data Collection Method**

Data used in this study were collected from the Ministry of National Education, Measurement, Evaluation, and Exam Services General Directorate. Sampling for the data used in this study was randomly performed by the Measurement, Evaluation, and Exam Services General Directorate and a data set including 10,000 students was given to the researcher. The researcher conducted the data cleaning process.

### **2.3. Process**

In this study, the construct of the data set was analyzed first. For this purpose, the data set was randomly divided into two parts, and while exploratory factor analysis (EFA) was applied to

one half, confirmatory factor analysis (CFA) was applied to the other. For EFA, it was first analyzed whether the data set met EFA assumptions. Accordingly, for multivariate normality, the skewness and kurtosis coefficients of Mardia (1970) were analyzed. Multivariate skewness and kurtosis coefficients showed that the data did not hold the assumption of multivariate normality ( $p < 0.01$ ). Therefore, the principal axis factoring method was adopted, which is stronger in terms of violation of the normality assumption (Costello & Osborne, 2005; Fabrigar, Wegener, MacCallum, & Strahan, 1999). When the adequacy of the sample size was analyzed, it was concluded that a sample of 4,886 was sufficient for the majority of researchers (Comrey, 1988; Floyd & Widaman, 1995; Gorsuch, 1974; Guadagnoli & Velicer, 1988; Kaiser & Rice, 1974; Leech, Barrett, & Morgan, 2015; Streiner, 1994). Additionally, it was observed that the KMO value was 0.95. Accordingly, it can be concluded that the sample was adequate for factor analysis and that an adequate number of items corresponded to each factor (Kaiser & Rice, 1974; Leech et al., 2015). The Bartlett test results, which analyzed whether the correlation matrix was different to the identity matrix, showed that it was ( $\chi^2(190) = 21775.9, p < 0.01$ ). On the other hand, Mahalanobis distances were calculated to analyze multivariate outliers. Among the 4,886 data in this sample, 145 Mahalanobis distances that provided significant results at the  $\alpha = 0.001$  level were deleted, and a data set of 4,741 people was obtained. For the multicollinearity assumption, the variance inflation factor (VIF), tolerance value (TV), and conditional index (CI) were analyzed since there should be no multicollinearity. It was observed that the tolerance value was larger than 0.01, the VIF value was smaller than 10, and the CI value was smaller than 30. Accordingly, it can be concluded that there was no multicollinearity problem (Kline, 2016; Tabachnik & Fidell, 2012).

Tetrachoric correlation matrix and principal axis factoring as factor extraction methods were applied to data divided into two for EFA. First, parallel analysis was conducted to determine the number of factors and the analysis proposed a unidimensional construct. On the other hand, when the scree plot and eigenvalues were analyzed, it was observed that only the eigenvalue of the first factor was larger than one. The unidimensional construct explained 47.96% of total variance. Therefore, it was decided that the test was unidimensional. When factor loadings were analyzed, it was observed that loadings changed between 0.44 and 0.79. Accordingly, it can be concluded that a unidimensional structure was defined as the result of EFA.

Confirmatory factor analysis (CFA) was applied to the second half of the data set. Again, assumptions of the analysis were investigated. As it was observed that multivariate normal distribution was not held, tetrachoric correlation matrix and weighted least squares means and variance adjusted (WLSMV) estimation methods were applied for CFA (Li, 2016). On the other hand, Mahalanobis distances were calculated to analyze multivariate outliers. From among the 4,887 data in this sample, 69 Mahalanobis distances that provided significant results at the  $\alpha = 0.001$  level were deleted, and a data set of 4,818 people was obtained. For multicollinearity assumption, the variance inflation factor (VIF), tolerance value (TV), and conditional index (CI) were analyzed since there should be no multicollinearity. It was observed that the TV was larger than 0.01, the VIF value was smaller than 10, and the CI value was smaller than 30. Accordingly, it can be concluded that there was no multicollinearity problem (Kline, 2016; Tabachnik & Fidell, 2012).

The results of CFA applied to the second half showed significant chi-square values ( $\chi^2(170) = 753.45, p < 0.01$ ). Accordingly, it can be said that the model data fit was not held. However, since this statistic has a tendency to be significant and high in large samples (Mueller, 1996), other fit statistics were examined. Accordingly, CFI and TLI values were observed as being 0.99. Additionally, all factor loadings had statistically significant t-values between 0.44 and 0.78. Error variances changed between 0.40 and 0.80. Based on these results, it can be concluded that data fitted with the unidimensional construct.



After determining that the data set was unidimensional, the total score, IRT ability estimation, and factor scores of each individual were calculated. The total scores of individuals were calculated from the total number of correct answers of individuals to each item. For ability estimation based on IRT, it was analyzed whether the data set held IRT assumptions (unidimensional, local independence and S-shape) (DeMars, 2010; Hambleton & Swaminathan, 1985; Lord, 1980). It was observed that unidimensional assumptions were held when the factor structure was investigated. Yen's (1984)  $Q_3$  statistic was used to determine whether the local independence assumption was held. For this purpose, the model data fit was analyzed and which unidimensional IRT model (1-, 2-, or 3-parameter logistic model) fitted the data set was investigated. Accordingly, log likelihood values were examined. When models were compared, it was found that the three-parameter logistic model (3PLM) fitted data better ( $\chi^2_{2PLM-3PLM}(20) = 1352.55, p < 0.01$ ). Item parameters were estimated based on 3PLM, the residual matrix was created with residuals of each item, and the correlation between them was analyzed. It was observed that correlations were not higher than the 0.20 threshold value. Accordingly, it can be concluded that the local independence assumption was held (DeMars, 2010). Whether item characteristic functions were S-shaped was analyzed by plotting the item characteristic curve and it was observed that they had an S-shape. Ability estimation of individuals was obtained with the expected a posteriori (EAP) method for IRT.

After obtaining total score and ability estimations, the Bartlett method was used to estimate factor scores. The Bartlett method was selected as it is unbiased when estimating real factor scores (Hershberger, 2005). After obtaining the total score, ability estimations, and factor scores of individuals, the relationship between three variables was analyzed using correlation analysis. Additionally, a scatter plot was used to visually represent the relationship between variables.

## 2.4. Data Analysis

In this study, to estimate EFA and factor scores, the psych package (Revelle, 2018) in R software (R Core Team, 2017) was used. *Mplus* software (Muthén & Muthén, 2012) was used for CFA. IRT parameter estimations were performed using the *irtoys* package (Partchev, 2016) in R, and the BILOG engine. The *sirt* package (Robitzsch, 2017) was used to test the local independence assumption. *ggplot2* (Wickham, 2016) in R software was used for plotting the graphs.

## 3. FINDINGS

In this study, the answer to the question “According to the relationship between total score, ability estimation, and factor score, can factor scores be used instead of ability estimation and total score?” was investigated. Accordingly, the relationship between total score, ability estimation, and factor scores was analyzed and presented in [Table 1](#).

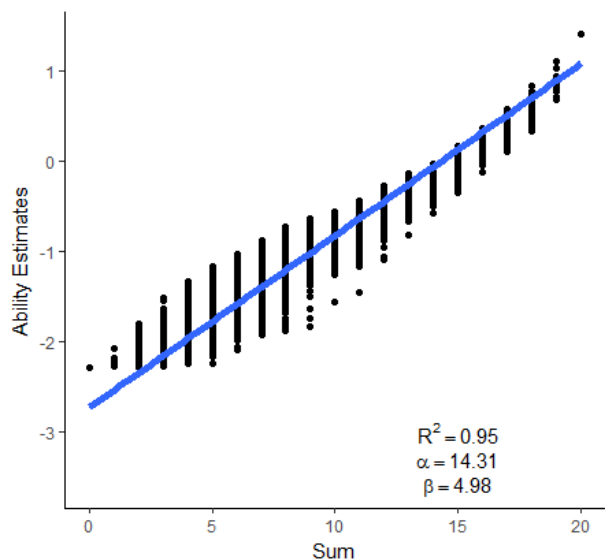
**Table 1.** The relationship between total score, ability estimation, and factor score

| Variables          | $\bar{X}$ | S    | Skewness | Kurtosis | Correlation |                    |              |
|--------------------|-----------|------|----------|----------|-------------|--------------------|--------------|
|                    |           |      |          |          | Total score | Ability Estimation | Factor Score |
| Total score        | 14.30     | 4.75 | -0.64    | -0.59    | 1           |                    |              |
| Ability Estimation | 0.00      | 0.93 | -0.20    | -0.71    | 0.975**     | 1                  |              |
| Factor Score       | 0.00      | 1.23 | -1.00    | 0.25     | 0.982**     | 0.960**            | 1            |

\*\*p<0.01

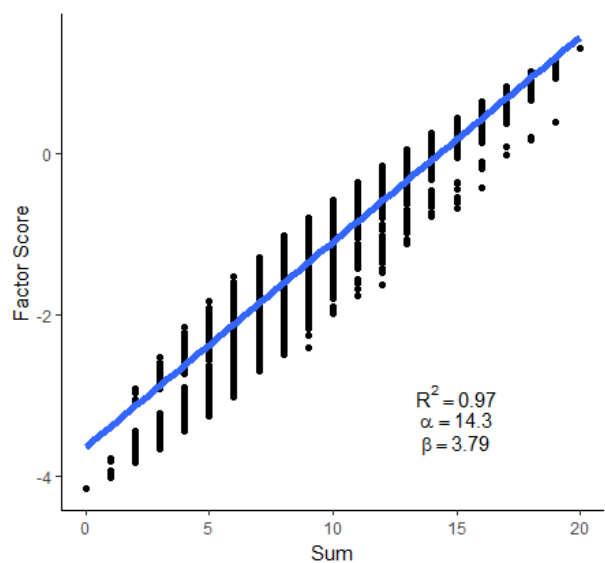
When [Table 1](#) was examined, a correlation between the total score, ability estimation, factor score, and descriptive statistics of variables was observed. Since the scales of total score, ability estimation, and factor score were different, it can be concluded that mean and standard deviation

values were different. When skewness and kurtosis values were examined, it was observed that skewness values were between -1.00 and -0.20 while kurtosis values were between -0.71 and 0.25. Accordingly, it can be concluded that variables have a univariate normal distribution (Byrne, 2016; Chou & Bentler, 1995; Curran, West, & Finch, 1996; Finney & DiStefano, 2013). Therefore, correlations between variables were calculated using the Pearson Product Moment (PPM) correlation coefficient. When correlations between variables were analyzed, it could be stated that there was a positive and high-level relationship. A scatter plot of the total score and ability estimation of individuals is presented in Figure 1.



**Figure 1.** Total score and ability estimate distribution.

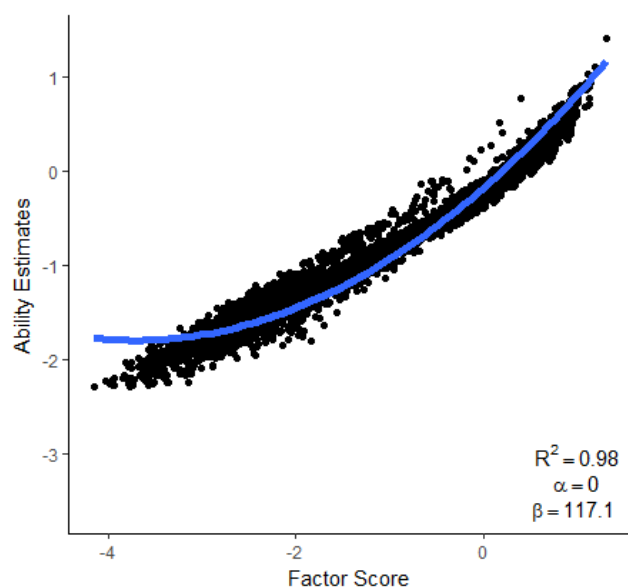
When Figure 1 is examined, the distribution of total score and ability estimation can be observed. The explained variance on the obtained linear regression equation was 95%. Accordingly, it can be stated that the variation in total score explained 95% of the variation in ability estimation. On the other hand, there was differentiation in the ability estimation for each total score category. For example, many students with a total score of 5 differed in ability estimation. A scatter plot of total scores and factor scores is presented in Figure 2.



**Figure 2.** Total score and factor score distribution.



When Figure 2 is examined, the distribution of total score and factor score can be observed. The explained variance on the obtained linear regression equation was 97%. Accordingly, it can be stated that the variation in the total score explained 97% of the variation in the factor score. On the other hand, there was differentiation in the factor score for each total score category. For example, many students with a sum of 10 differed in their factor score. A scatter plot of ability estimation and factor score is presented in Figure 3.



**Figure 3.** Factor score and ability estimates distribution.

When Figure 3 is examined, it can be concluded that the relationship between total score, ability estimation, and factor score was nonlinear. While the linear relationship presented in Table 1 explained 92.16% ( $0.962=0.921$ ) of the variance, when the relationship between the two variables was considered as quadratic, 98% of the variance was explained. Accordingly, it can be stated that the relationship between the factor score and ability estimation was high.

#### 4. DISCUSSION AND CONCLUSION

In the results of this study, a positive, statistically significant, and strong relationship between ability estimation and total score was observed. This finding is similar to findings of other studies in the literature (Bulut, 2018; Çelen & Aybek, 2013; Fava & Velicer, 1992; Grice, 2001; Macdonald & Paunonen, 2002; Progar & Sočan, 2008; Velicer, 1976). Based on this result, it can be said that the factor score, ability estimates, and total score obtained from the high-stakes achievement test are strongly related to each other. Due to the high relationship between the two variables, it is claimed that these variables can be used interchangeably. Tabachnik and Fidell (2012) stated that the relationship between the variables was 0.90 or higher, one of the variables was redundant, or this variable is a combination of other variables. From this perspective, a 0.975 value of correlation between variables showed that instead of total score, ability estimation can be used. When ability estimation was used, in contrast to total score, individual ability estimations in one category of sum differed. In this case, IRT influences ability estimation by using item parameters when estimating ability.

When the relationship between total score and factor scores was analyzed, it could be concluded that a similar high, positive, and statistically significant relationship was present for ability estimation. Similarly, in ability estimation, individuals with the same total score had different factor scores.

It was observed that the relationship between factor score and ability estimation was nonlinear. The relationship between these two variables signified a quadratic function and, in this case, it was observed that explained variance was extremely high. In other words, it could be claimed that factor score and ability estimation can be used interchangeably.

Total score is practical in terms of calculation and interpretation. On the other hand, if ability estimation based on IRT is used, real abilities of individuals can be estimated. However, large samples are needed to hold IRT assumptions. DeMars (2010) stated that as the number of parameters increases, and as the ability distribution of the group moves away from normality, the sample size must grow. When the number of items is 20 and discrimination parameters are high, a sample size of at least 500 is needed. If the pseudo chance parameter is estimated, the sample size should be at least 2000 (DeMars, 2010). When these limitations of IRT is considered, it can be expressed that factor scores are more efficient. In many cases, factor scores can be estimated from a sample size of 250. Floyd and Widaman (1995) stated that there should be four or five individuals per item and the sample size should be as large as possible. Streiner (1994) suggested that each item should contain five individuals and the sample should not be smaller than 100. If the sample size should be smaller than 100, 10 individuals should be sampled for each item. Gorsuch (1974) suggested that each item should contain five individuals and the sample size should not be smaller than 200. Guadagnoli and Velicer (1988) stated that these calculations were baseless and EFA can be applied a sample size smaller than 50 when factor loadings are 0.80 or higher. Comrey (1988) stated that if the number of items did not exceed 40, a sample size of 200 individuals would be sufficient. When all these recommendations of these researchers were considered, it could be concluded that EFA needs a smaller sample size than IRT. Additionally, mainly EFA is conducted for the unidimensionality assumption of the IRT assumption evaluation stage. Therefore, it could be concluded that the practicability of factor score estimation is better than IRT ability estimation.

Based on the findings of this study, since it was observed that total score, ability estimation, and factor score can be used interchangeably, factor score is recommended. Because factor score requires smaller sample size, discriminates individuals better than total score, and produces very close result with IRT ability estimation. On the other hand, Erkuş (2014) suggested item weighting via factor loadings. He recommends that individual responses to items and that item factor loading multiplication (if individual score is 1 and factor loading is 0.48 the item score is  $1 \times 0.48 = 0.48$  for that individual) for calculate total score. But factor scores are also calculated with factor loadings as well as other elements of factor analysis such as eigenvalues, communalities, and error variance. In this sense, using factor scores can be suggested. After estimating factor scores, linear transformation or T points can be used for reporting and results can be interpreted more easily.

The current study is limited by unidimensional constructs. Therefore, multidimensional constructs may be studied in future studies. On the other hand, since the data set of the current study was extremely large, a smaller sample size can be investigated in another study. As a result of this study, since it is thought that factor scores will make a positive contribution to the validity of decisions regarding individuals, the use of factor scores is suggested.

## ORCID

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

## 5. REFERENCES

- Akyıldız, M., & Şahin, M. D. (2017). Açıköğretimde kullanılan sınavlardan Klasik Test Kuramına ve Madde Tepki Kuramına göre elde edilen yetenek ölçülerinin karşılaştırılması. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, 3(4), 141–159.

- Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability* (Vol. 5, pp. 111–150). Berkeley: University of California Press.
- Bulut, G. (2018). Açık ve uzaktan öğrenmede şans başarısı : Klasik Test Kuramı (KTK) ve Madde Tepki Kurama (MTK) temelinde karşılaştırmalı bir analiz. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, 4(1), 78–93.
- Büyüköztürk, Ş., Kılıç-Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2013). *Bilimsel araştırma yöntemleri* (14. Baskı). Ankara: Pegem Akademi.
- Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). Routledge.
- Çakıcı-Eser, D. (2013). PISA 2009 okuma testinden elde edilen iki kategorili verilerin BILOG programı ile incelenmesi. *Eğitim ve Öğretim Araştırmaları Dergisi*, 2(4), 135–144.
- Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of Classical Test Theory and Item Response Theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, 36(5), 648–662. <https://doi.org/10.1016/j.clinthera.2014.04.006>
- Çelen, Ü., & Aybek, E. C. (2013). Öğrenci başarısının öğretmen yapımı bir testle klasik test kuramı ve madde tepki kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 4(2), 64–75. Retrieved from <http://dergipark.ulakbim.gov.tr/epod/article/view/5000045503>
- Chou, C. P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: SAGE.
- Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, 56(5), 754–761. <https://doi.org/10.1037/0022-006X.56.5.754>
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). London: Lawrence Erlbaum Associates, Inc. <https://doi.org/10.1017/CBO9781107415324.004>
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10(7), 27–29. <https://doi.org/10.1.1.110.9154>
- Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed Methods approaches* (4. Edition). Thousand Oaks, CA: Sage Publications.
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1(1), 16–29. <https://doi.org/10.1037/1082-989X.1.1.16>
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press. <https://doi.org/10.1073/pnas.0703993104>
- DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.
- DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, 14(20), 1–11.
- Erkuş, A. (2014). *Psikolojide ölçme ve ölçek geliştirme-I: Temel kavramlar ve işlemler* (2. Baskı). Ankara: Pegem Akademi.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272–299. <https://doi.org/10.1037/1082-989X.4.3.272>
- Fava, J. L., & Velicer, W. F. (1992). An empirical comparison of factor, image, component, and scale scores. *Multivariate Behavioral Research*, 27(3), 301–322.

- [https://doi.org/10.1207/s15327906mbr2703\\_1](https://doi.org/10.1207/s15327906mbr2703_1)
- 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). Charlotte, NC: IAP.
- Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7(3), 286–299.
- Fraenkel, J. R., Wallen, N. E., & Huyn, H. H. (2012). *How to design and evaluate research in education* (8. Edition). New York: McGraw-Hill.
- Gorsuch, R. L. (1974). *Factor analysis* (1st ed.). Toronto: W. B. Saunders Company.
- Green, B. F. (1976). On the factor score controversy. *Psychometrika*, 41(2), 263–266. <https://doi.org/10.1007/BF02291843>
- Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6(4), 430–450. <https://doi.org/10.1037//1082-989X.6.4.430>
- Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103(2), 265–275.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. New York: Springer Science & Business Media, LLC.
- Hershberger, S. L. (2005). Factor score estimation. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 2, pp. 636–644). Chichester, UK, UK: John Wiley & Sons, Ltd. <https://doi.org/10.1002/0470013192.bsa726>
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185. <https://doi.org/10.1007/BF02289447>
- İlhan, M. (2016). Açık uçlu sorularda yapılan ölçmelerde Klasik Test Kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe University Journal of Education*, 31(2), 346– 368. <https://doi.org/10.16986/HUJE.2016015182>
- Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, 34(1), 111–117. <https://doi.org/10.1177/001316447403400115>
- Kline, R. B. (2016). *Principle and practice of structural equation modelling* (4th ed.). New York, NY: The Guilford Press.
- Leech, N. L., Barrett, K. C., & Morgan, G. A. (2015). *IBM SPSS for intermediate statistics* (5. Baskı). East Sussex: Routledge.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. *Journal of Chemical Information and Modeling* (Vol. 53). New Jersey: Lawrence Erlbaum Associates, Inc. <https://doi.org/10.1017/CBO9781107415324.004>
- Macdonald, P., & Paunonen, S. V. (2002). A monte carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, 62(6), 921– 943. <https://doi.org/10.1177/0013164402238082>
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519–530.
- Mueller, R. O. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EQS. Design* (Vol. 102). New York: Springer Science & Business Media, LLC. <https://doi.org/10.1016/j.peva.2007.06.006>
- Muthén, L. K., & Muthén, B. O. (2012). Mplus statistical modeling software: Release 7.0. *Los Angeles, CA: Muthén & Muthén*.
- Partchev, I. (2016). irtoys: A collection of functions related to item response theory (IRT). Retrieved from <https://cran.r-project.org/package=irtoys>

- Price, L. R. (2017). *Psychometric methods: Theory and practice*. New York, NY: The Guilford Press.
- Progar, S., & Sočan, G. (2008). An empirical comparison of Item Response Theory and Classical Test Theory. *Horizons of Psychology*, 17(3), 5–24.
- R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.r-project.org/>.
- Revelle, W. (2018). psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois. Retrieved from <https://cran.r-project.org/package=psych>
- Robitzsch, A. (2017). sirt: Supplementary item response theory models. Retrieved from <https://cran.r-project.org/package=sirt>
- Stage, C. (1998a). *A comparison between item analysis based on item response theory and classical test theory. A study of the SweSAT Subtest ERC. Educational Measurement*. Retrieved from [http://www.sprak.umu.se/digitalAssets/59/59551\\_enr2998sec.pdf](http://www.sprak.umu.se/digitalAssets/59/59551_enr2998sec.pdf)
- Stage, C. (1998b). *A comparison between item analysis based on item response theory and classical test theory. A study of the SweSAT Subtest WORD. Educational Measurement*. Retrieved from [http://www.sprak.umu.se/digitalAssets/59/59551\\_enr2998sec.pdf](http://www.sprak.umu.se/digitalAssets/59/59551_enr2998sec.pdf)
- Streiner, D. L. (1994). Figuring out factors: The use and misuse of factor analysis. *Canadian Journal of Psychiatry*, 39(3), 135–140.
- Tabachnik, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6. ed.). Boston: Pearson.
- Velicer, W. F. (1976). The relation between factor score estimates, image scores, and principal component scores. *Educational and Psychological Measurement*, 36(1), 149–159. <https://doi.org/10.1177/001316447603600114>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <http://ggplot2.org>
- Williams, J. S. (1978). A definition for the common-factor analysis model and the elimination of problems of factor score indeterminacy. *Psychometrika*, 43(3), 293–306. <https://doi.org/10.1007/BF02293640>
- Xu, T., & Stone, C. A. (2012). Using IRT trait estimates versus summated scores in predicting outcomes. *Educational and Psychological Measurement*, 72(3), 453–468. <https://doi.org/10.1177/0013164411419846>
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125–145. <https://doi.org/10.1177/014662168400800201>



## Impact of Emotional Literacy Training on Students' Emotional Intelligence Performance in Primary Schools

Kerem Coskun <sup>1,\*</sup>, Yucel Oksuz <sup>2</sup>

<sup>1</sup> Department of Primary Education, Artvin Coruh University, Artvin, Turkey

<sup>2</sup> Department of Educational Sciences, Ondokuz Mayis University, Samsun, Turkey

### ARTICLE HISTORY

Received: 31 March 2018

Revised: 01 December 2018

Accepted: 12 December 2018

### KEYWORDS

Emotional Intelligence,  
Emotional Literacy,  
Primary School Students,  
Social Development,  
Emotional Development

**Abstract:** The present study seeks to reveal the impact Emotional Literacy Training (ELT) that lasted for two months, on students' emotional intelligence performance. The study was designed as a quasi-experimental research. The experimental group consisted of 16 students, while 12 students were assigned to control group. Data in pre-test and post-test were collected through the Ten Years Emotional Intelligence Scale (TYEIS) developed by Coskun, Oksuz and Yilmaz (2017). Data were analysed through the paired sample t-test and independent sample t-test. Findings of the study indicated that ELT significantly increased experimental group students' emotional intelligence performance and this significant increase remained permanently. Results were discussed according to Experiential Learning and Radical Behaviourism and in the light of relevant literature; several implications were developed for teachers, other school staffs and researchers.

## 1. INTRODUCTION

Emotions have a critical impact on mental health, morality, and spirituality, learning and cognitive functions. Therefore, social and emotional skills are key components of the educational process to sustain children's developmental process, conduct an effective instruction. The fact that emotional skills are so crucial that requires conceptualization and systematic instruction.

Emotional intelligence is one of the conceptualizations related to social-emotional skills. Emotional intelligence can be described as a construct which consists of self-awareness, appropriate explanation of emotions, self-regulation, motivation, and establishing positive relationships with others. There are three emotional intelligence models. Those are the ability model, the mixed models, and the trait model (Matthews, 2006).

---

CONTACT: Kerem Coskun ✉ [keremcoskun@artvin.edu.tr](mailto:keremcoskun@artvin.edu.tr) 📍 Department of Primary Education, Artvin Coruh University, Artvin, Turkey

ISSN-e: 2148-7456 / © IJATE 2019



Goleman (1998) proposed an emotional intelligence model. It has five sub-dimensions as self-awareness, self-management, empathy, motivation, and social skills. Self-awareness is the skill to accurately recognize and label emotions in self. Self-management skill is the capacity to effectively cope with emotions. Empathy is a very crucial skill in the model because empathy has a function to establish and prolong constructive relationships with others. Empathy can be described as a social skill to recognize emotions in others understand and respond them. Therefore, Goleman (1998) sees empathy as social radar. Motivation is the skill which includes motivation drive, commitment, initiation, and optimism (Boyatzis, Goleman, & Rhee, 2000). Social skills are the skills which help the individual to establish cooperative and positive relationships with others and sustain these relationships. Social skills consist of influence, conflict management, leadership, change catalyst, building bonds, collaboration and cooperation, and team capabilities (Boyatzis, Goleman & Rhee, 2000).

Emotional literacy, developed by Steiner (1979) is another term about the emotional and social world. Steiner (1979) defined emotional literacy as a construct which includes accurate recognition of emotion in self, empathy, suitable express of emotions to others, and emotion management. Steiner (1979) divides the emotional literacy into five skills. Those are knowing emotions, possession of a sense of empathy, management of emotions, the resiliency of emotional damage, and combining those skills.

Faupel (2003) practically addressed emotional literacy and developed an emotional literacy model. Faupel Emotional Literacy Model consists of self-awareness, self-regulation, motivation, social competence, and social skills. Self-awareness is the skill which helps to label and name emotions. Self-regulation emphasizes learning to control emotions and modify behaviours. Motivation is another component of the emotional literacy. Motivation enables students to determine a goal and initiate to act. In Faupel's Emotional Literacy Model, social competence means empathy as a skill to understand others' emotions needs, and concerns. Social skills include conflict resolution, influence, communication, and leadership, change catalyst, building bonds, team capabilities, collaboration and cooperation.

There may be confusion between emotional intelligence and emotional literacy. Moreover, both of the concepts may be used interchangeably each other. Emotional intelligence is preferred to use in the USA, educators in the UK opt for the emotional literacy. There are differences between the two concepts, even though emotional literacy and emotional intelligence have similar characteristics. The concept of intelligence means the individual capacity to progress either cognitive information or socio-emotional information, where the term of literacy emphasizes possession of linguistic skills and strategy about how linguistic skills are put into practice in order to exchange ideas with others and overcome language barriers in daily living. In other words, literacy centres on possession of skills and strategic insight about their practice while intelligence makes an emphasis potential capacity to progress information (Matthews, 2006). Emotional intelligence has individualistic characteristics about a progression of socio-emotional information. Emotional literacy can be considered as an instruction and strategy in which student is taught socio-emotional skills (Orbach, 1998; Tew, 2007; Pratt, 2009). Social and emotional Aspect of Learning (SEAL) conducted in England, Promoting alternative Thinking Strategies (PATHS) and Collaborative for Academic, Social and Emotional Learning (CASEL) are instructional curriculums whose purpose is to teach socio-emotional skills in schools. Therefore, those curriculums are known as emotional literacy programs (Goleman; 1995; Park, 1999; Perry, Lennie & Humphrey, 2008; Burman, 2009; Hallam, 2009; Pratt, 2009; Flynn, 2010; Gillum, 2010). The more educational goals about social and emotional learning are included, the more classroom time is allocated to teach social and emotional skills contemporary instructional curriculums.

A few number of the research indicated that increase in emotional intelligence performance of students has positive outcomes in academic achievement, more competency to cope with depression, better adjustment, and social support (Qualter, Whiteley, Hutchinson & Pope, 2007; Hallam, 2009; Ferrando et al., 2010; Di Fabio & Kenny, 2011; Jellesma, Rieffe, Terwogt & Westenberg, 2011; Rivers et al., 2012). In the present study, it was expected that teaching emotional intelligence skills from Goleman's emotional intelligence model through emotional literacy training designated by the researchers, may reveal beneficial outcome for children who are at the end of primary school period. Furthermore, there is a close relationship between emotional intelligence and spirituality (Paek, 2006). Improving emotional intelligence through emotional literacy activities can lead to better spirituality, well-being, and social functioning. Therefore, instructional emotional literacy activities may trigger the development of spirituality skills among primary school children (Emmons, 2000).

**Purpose of the study:** Schools aim to develop social-emotional skills as well as cognitive skills of students. Emotional intelligence is a concept that should be fostered in schools in the context of social-emotional skills. Development of emotional intelligence skills allows students to recognize and express their emotions in a constructive way, cope with stringent emotions, develop sense of empathy, and establish positive relationships with others. However, development of emotional intelligence skills entails systematic and planned instruction.

Concept of emotional intelligence defines necessary skills for social-emotional adjustment but it does not offer any explanation about how to teach and foster the skills. Emotional literacy offers an explanation about how to teach. In the UK and the USA the SEAL, the CASEL, and the PATHS are emotional literacy programs that foster emotional intelligence skills in schools from primary schools to high schools (Goleman; 1995; Park, 1999; Perry, Lennie & Humphrey, 2008; Burman, 2009; Hallam, 2009; Pratt, 2009; Flynn, 2010; Gillum, 2010).

In Turkey, social-emotional skills are dealt with in life knowledge and social studies courses. However, emotional intelligence is taught within limited duration and there is no systematic and planned teaching approach. Although there are a few studies that develop social skills, emotional intelligence in Turkey, but those studies were conducted with research participants who had been isolated from their previously established social interactions (Arda & Ocak, 2012; Saltalı, 2011; Ulutaş, 2005; Yaşarsoy, 2006). Purpose of emotional literacy is to develop social-emotional skills of students based on their social interactions and without isolating from classroom environments. In the present study, a specific emotional literacy training program was developed and conducted with primary school participants without isolating from their classroom environments, and its impact was explored. As a result of the present study, it was sought out modelling an emotional literacy training and offering practical implications for Turkish primary education system.

## 2. METHOD

**Design of the Research:** The present study was designed in experimental research, one of the quantitative research traditions, because of the fact that the present study aims to reveal the impact of the ELT, seeks causation between the ELT and scores of the emotional intelligence performance of the participant primary school children. In experimental researches, the impact of an independent variable upon dependent variable is sought to reveal. For this reason, the experimental group was manipulated by ELT, independent variable of the study, while the control group received no manipulation. Participant students were not randomly and individually assigned to the groups because of the fact that the emotional literacy training depends on previously established interaction among the students. The present study was designed in quasi experimental research due to the impossibility of random assignment to the groups (Cohen, Manion, & Morrison, 2000; Frankel, Wallen, & Hyun, 2012).

**Selection of the Research Participants:** Convenience sampling was used because of the fact that random selection of the participant was impossible. Emotional literacy depends on previously established social interaction among the participants. This notion prevented random selection and creation of a new study group. Research participants were selected according to voluntary participation of primary school children and their teachers, classroom size, ages of the participants. Classroom size was as crucial as voluntary participation, because of the fact that ELT depends on a careful focus on interaction, active participation, discussion instructional experience in depth, and sharing thoughts and emotions with others. Therefore, it was decided that a maximum number of the members in both of the groups should not exceed 20. On the other hand, measurement of emotional intelligence was carried out through self-report, and self-report depends on sincere and accurate response without any bias. Primary school children who are 10 years old are more adept to sincerely and accurately evaluate the items on the TYEIS. Therefore, it was decided that primary school children at the age of 10, would be included in the study. Before consulting primary school children and their teachers, necessary official permission was taken from local education authorities in Turkey. Upon receiving official permission, the researchers visited primary schools, met teachers and primary school children, explained what would be conducted. Two primary school teachers and their children accepted to participate in the study. The groups were matched through their scores from the TYEIS developed by Coskun, Oksuz & Yilmaz (2017). As a result of the application of the TYEIS as a pre-test, 16 students whose age is 10, were assigned to the experiment group while the control group consisted of 12 students whose age is 10 years. In the experiment group, 7 of primary school children were female while 9 of them were male. In the control group, 6 of the primary school children were female whereas 6 of them were male.

In order to test whether the groups were equal to each other and parametric or non-parametric test would be used, normality test was conducted. Shapiro-Wilk test was used in the normality test because of fewer students in the groups than 30 (Shapiro-Wilk, 1965). Results of the normality test were indicated in Table 1.

**Table 1.** Result of the Normality Test

| Measurement                | Groups     | n  | Shapiro-Wilk (S-W) | $\bar{X}$ | Df | Sd   |
|----------------------------|------------|----|--------------------|-----------|----|------|
| Total Score from the TYEIS | Experiment | 16 | .11                | 22.50     | 16 | 1.36 |
|                            | Control    | 12 | .98                | 23.08     | 12 | 3.06 |

Normality test analysis indicated that the data has a normal distribution. As a result of the normal distribution, a comparison between both of the groups was made through the independent t-test (Field, 2009).

**Table 2.** Results of Independent T-Test

| Measurement                | Groups     | n  | $\bar{X}$ | Ss   | Sd | t    | p   |
|----------------------------|------------|----|-----------|------|----|------|-----|
| Total Score from the TYEIS | Experiment | 16 | 22.50     | 1.36 | 26 | -.68 | .49 |
|                            | Control    | 12 | 23.08     | 3.02 |    |      |     |

\* $p=0.05$

Results of independent t-test revealed that there is no significant difference between the groups in total scores of the TYEIS ( $t_{(26)} = -.68, p > 0.05$ ). Therefore, it was concluded that the experiment group and the control group were equal to each other in pre-test measurement.

**Components of The Emotional Literacy Training Activities:** ELT Activities were designed for self-awareness, self-regulation, motivation, empathy, and social skills. In the self-awareness activities, the participant children's emotional vocabulary was expanded by displaying human faces, circle time discussions, work-sheets to which they responded the questions about environment and emotions. In ELT Activities for self-regulation skill, the participant children learnt to how to manage emotions and modify behaviours through criticise, competition, puzzle games. ELT Activities for motivation skill helped the participant children handle with excitement so as to be motivated through classroom-based competitions and follow-up discussions. ELT Activities for empathy skill fostered empathy among the participant children by pairing each other, stories. In the social skill activities, participant children constituted groups and cooperated in order to design a poster, contemplated on the events on which one of them did a favour for another friend, and produced "*the Sharing Tree*". Furthermore, at the end of each of the activities, each participant children stated what they had felt, thought, behaved.

**The Process:** There is a close similarity between Goleman Emotional Intelligence model and Faupel Emotional Literacy Model in terms of categorizing social and emotional skills so the two models were integrated into the present study. The researchers investigated Goleman Emotional Intelligence Model and they decided which social-emotional skills would be taught during ELT. On the other hand, Faupel Emotional Literacy Model inspired the researchers how the social-emotional skills would be taught. In other words, the social-emotional skills from Goleman Emotional Intelligence Model set instructional goals, while emotional literacy model by Faupel (2003) functioned as instructional ways how to teach. 16 social and emotional skills were determined as instructional aims; 18 instructional activities were designed to teach those skills. A pilot study was conducted in the 2013-2014 instructional year between March and June. After pilot study, it was concluded that 18 instructional activities were decreased to 16 instructional activities. As a result of the pilot study, duration of the activities revised and more classroom time was allocated to students to state their emotions and experiences which they underwent in the activities.

Before the ELT, the TYEIS were applied on both of the groups as the pre-test. The experimental group received the ELT during 8 weeks. ELT lasted for 31-course hours during 8 weeks. ELT was administered by the researcher. The control group received no treatment so previously planned and ordinary classroom activities were applied to the control group. After ELT had finished, participant students from both the experimental groups and the control group took the TYEIS as post-test. The TYEIS were again given to the participant students in both of the groups as follow-up test.

**The Instruments:** Emotional intelligence performance of the participant students was assessed through the TYEIS developed by Coskun, Oksuz & Yilmaz (2017). The TYEIS measures social and emotional skills in the Goleman emotional intelligence model. It consists of ten items with one dimension. In the TYEIS the items have three response choices as "*not true*", "*somewhat true*", and "*completely true*". All of the items are so negative that they were reversely graded. The highest point is 30 and the lowest point is 10 in the TYEIS. Its reliability coefficient is .89 and it has also good model fitting indices (RMSEA= .06, CFI=.97, IFI=.97, RFI= .93, GFI=.95, AGFI=.94, NFI= .95, SRMR= .03).

### 3. FINDINGS

Levene test was carried out in so as to test homogeneity assumptions. Results of Levene test were indicated in [Table 3](#).

Result of Levene test indicated that homogeneity assumption was confirmed for both the scores of the experiment group and that of the control group in the TYEIS. Homogenous variance in both of the groups' scores enabled using Analysis of Covariance (ANCOVA) to compare

emotional intelligence performance of the experiment group to emotional intelligence performance of the control group. Moreover, ANCOVA helps to reduce the error variance and allows assessing more precisely the impact of ELA, independent variable of the study (Field, 2009; Tabachnik & Fidell, 2012). Therefore, due to homogenous variances, reduction in the error variance, and more precise assessment of the ELA, ANCOVA was used to identify impact of the ELA. ANCOVA results were displayed in [Table 4](#).

**Table 3. Homogeneity Test Through Levene Test**

| Measurement | F    | Sd <sub>1</sub> | Sd <sub>2</sub> | P    |
|-------------|------|-----------------|-----------------|------|
| TYEIS       | 7.10 | 1               | 26              | .073 |

P= .05

**Table 4. ANCOVA Results**

| Factor                    | SS     | Df | MS     | F     | P    | $\eta^2$ |
|---------------------------|--------|----|--------|-------|------|----------|
| Corrected Model           | 228.47 | 2  | 76.13  | 7.71  | .001 | .49      |
| Intercept                 | 26.54  | 1  | 26.54  | 2.68  | .114 | .10      |
| TYEIS <sub>pre-test</sub> | 20.95  | 1  | 20.95  | 2.12  | .158 | .08      |
| Group                     | 186.35 | 1  | 186.33 | 18.86 | .001 | .44      |
| Error                     | 237.01 | 25 | 9.87   |       |      |          |
| Total                     | 594.00 | 28 |        |       |      |          |
| Corrected Total           | 465.42 | 27 |        |       |      |          |

Findings from [Table 4](#) indicated that there is a significant difference between the experimental group students' total score of the TYEIS and that of the control group student and the ELT significantly increased experimental group students' emotional intelligence performance ( $F_{(1, 25)} = 18.86, p < .05, \eta^2 = .44$ ).

Follow-up test was administered on both of the groups two months later. In order to test homogeneity assumption Levene test was conducted. Levene test results about the follow-up test were shown in [Table 5](#).

**Table 5. Homogeneity Test Through Levene Test**

| Measurement | F    | Sd <sub>1</sub> | Sd <sub>2</sub> | P   |
|-------------|------|-----------------|-----------------|-----|
| TYEIS       | 7.07 | 1               | 26              | .06 |

P= .05

Findings related to Levene test of the follow-up test indicated that variances in the experimental group and the control group are homogeneous. As result of the Levene test, it was decided that data are suitable to conduct ANCOVA in order to detect whether significant difference between the groups exists and the impact of the ELT is permanent (Field, 2009; Tabachnik & Fidell, 2012). Findings of ANCOVA were indicated in [Table 6](#).

Findings of ANCOVA indicated that there is a significant difference between experimental group students' total score and the control group students' total score in favour of the experimental group students ( $F_{(1, 25)} = 19.69, p < .05, \eta^2 = .44$ ). The significant difference emerged by the ELT in the post-test, continues to exist in the follow-up test.



**Table 6.** ANCOVA Results Follow-Up About Follow-Test

| Factor                    | SS     | Df | MS     | F     | P   | $\eta^2$ |
|---------------------------|--------|----|--------|-------|-----|----------|
| Corrected Model           | 228.47 | 2  | 114.20 | 12.09 | .00 | .49      |
| Intercept                 | 29.45  | 1  | 29.43  | 3.10  | .09 | .11      |
| TYEIS <sub>pre-test</sub> | 20.97  | 1  | 20.97  | 2.21  | .14 | .08      |
| Group                     | 186.68 | 1  | 186.68 | 19.69 | .00 | .44      |
| Error                     | 237.02 | 25 | 9.48   |       |     |          |
| Total                     | 594.00 | 28 |        |       |     |          |
| Corrected Total           | 465.42 | 27 |        |       |     |          |

#### 4. DISCUSSION

In the study, it was observed that ELT increased significantly experimental group's emotional intelligence performance while control group student's emotional intelligence performance did increase. Results support the notion that students' emotional intelligence performance can be developed through interventional programs and instructional activities. Results of the study are supported by several findings of the research in the literature ((Bredacs, 2010; Brown, 2003; Coppock, 2007; Di Fabio & Kenny, 2011; Dolev, 2012; Domitrovich, Cortes & Greenberg, 2007; Dulewicz & Higgs, 2004; Gillum, 2010; Haddon, Goodman, Park & Crick, 2005; Perry et al., 2008; Hallam, 2009; Hamre, Pianta, Mashburn & Downer, 2012; Kelly, Longbottom, Pots & Williamson, 2004; Lu & Buchanan, 2014; O'Hara, 2011; Zeidner, Roberts & Matthews, 2002). 10 years old children's emotional intelligence performance was developed through the ELT in the study. When the fact that emotional intelligence development is fixed at the age of 17, is taken into consideration, development emotional intelligence of children, whose age is 10, through ELT is important. Because 10 years old is a period in which students experience transition between primary school and secondary school, late childhood and puberty. Therefore, 10 years old children are vulnerable to the factors disrupting social-emotional development (Keefer, Holden, & Parker, 2013). Systematic and planned development of emotional intelligence performance of children during primary school period makes them more competent to cope with problems emerging in transitions between primary school and secondary school, late childhood and puberty.

Results of the study indicated that ELT increased significantly emotional intelligence performance of the experimental group students whereas the control group students' emotional intelligence performance did not increase. This significant difference is expected to stem from the ELT that is the independent variable of the study. During ELT, which lasted for eight weeks the experimental group's students were allowed to contemplate on their emotions, categorize, label, and express their emotions, care others' emotions, develop an awareness of social skills through previously established interactions each other. If emotional intelligence is an expression of social-emotional competency in long-term memory, possession of emotional intelligence skills can be developed through appropriate experiences, which are organized in classrooms. Furthermore, emotional intelligence skills can be reinforced and stabilized by appropriate experiences. Either long-term or short-terms achievements in experiences about emotional intelligence skills can pave the way of good and strong personality (Zeidner et. al., 2002). In ELT process, students underwent appropriate experiences through the classroom activities based on their previously interactions with each other. These appropriate and successful social-emotional experiences may be the first step in acquiring the good and strong personality. Therefore, teachers can design and employ emotional-literacy activities to compose positive classroom settings and decrease behavioural problems among students. On the other hand, school staffs, school workers aiming to increase socio-emotional learning develop school-



based emotional literacy activities and carry out school-wide to establish better school climate and overcome behavioural problems.

Results of the study can be explained through Experiential Learning Theory developed by Kolb (1984). Experiential Learning Theory deals with learning as a process in which knowledge is acquired through experiential transformation. The learner must think on and be open to experiences in depth, conceptualize experience by comparing other experiences in terms of similarities and differences, develop problem-solving and decision-making skills. Questions such as “*What did you realize?*”, “*Why did it happen?*” are important to allow learners to contemplate on their experiences in depth. Contemplation on experience and its analysis enable learners to focus on experiences. Therefore, emotions are an inseparable part of learning (Kolb, 1984; Jacobson & Ruddy, 2004;). In the ELT process the experimental group students were allowed to focus and think on their existing interactions with each other, left directly to social-emotional experiences, on the other hand, self-awareness, self-management, motivation, empathy, and social-relational skills were made more tangible by asking those directive questions. Therefore, emotional intelligence performance of the experimental group students was thought to be developed through experiential transformation.

Results about the follow-up of impact of ELT demonstrated that ELT increased permanently emotional intelligence performance of the experimental group students. Results of the study were supported by several research results in the literature (Greenberg & Kushe, 1998; Jones, Brown, & Aber, 2011; Reddy & Richardson, 2008). Permanent impact of ELT can be explained through radical behaviourism developed by Skinner (1984). Skinner (1984) claimed in his theory that environment in which a specific behaviour occurs, and results of a behaviour are important in learning defined as permanent behaviour change. Skinner (1984) also argued that a specific response, which generates successful outcomes, is repeated in similar environments and stimulus. Therefore, learning is dependent on outcome and environment. In ELT process, behaviours of the students that are appropriate for emotional intelligence were reinforced and they were allowed to realize that those appropriate behaviours produced successfully and desired outcomes and inappropriate behaviours did not work in positive outcomes and led to failures. Thus the experimental group students who were aware of those appropriate behaviours for emotional intelligence, brought about positive and successful outcomes, may have made those behaviours permanent by repeating in the classroom and school environments.

Emmons (2000) stated that there is an interaction between emotional intelligence and spirituality, spiritual skills can be developed instructional activities that are similar to the emotional intelligence training. As a consequence, fostering emotional intelligence among the participant children can result in improvement of spiritual skills. Therefore, it can be said that developing emotional intelligence performance through the emotional literacy activities has a multifaceted outcome for primary school children.

#### ***Limitations of the Study:***

- The ELT was designed for the 4<sup>th</sup> grade primary school children. The ELT can be designed for 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> grade primary school children, their impact can be monitored and revealed.
- The present study was designed in quasi experimental research with control group. Impact of the ELT can be explored through more robust experimental design in future research.
- The present study can be replicated with a larger population in future research.
- Impact of the ELT on 10 years old primary school children was investigated in the context of emotional intelligence. Impact of the ELT can be addressed in terms of different social-emotional skills concept such as social-skills, empathy, pro-social behaviour, emotional skills, emotion regulation, through different instruments.

### **Practical Implications**

- Developing emotional literacy training for all mandatory educational level allows social and emotional learning to be systematically dealt with from primary school to high school.
- In the present study impact of the ELT was investigated through experimental research that is one of the quantitative research methods. Impact of the ELT can be examined through phenomenological study, which is one of the qualitative research methods.
- Planned and systematic emotional literacy activities can develop students' socio-relational skills.
- Emotional literacy activities can improve students' skills in recognition and expression of their emotions.
- Teachers or other school staffs can increase empathic skills of students through emotional literacy activities.
- Results of the study revealed that the ELT increased participant student's emotional intelligence performance. In-service training can be developed to train teachers about how to design and conduct emotional literacy activities.
- Emotional literacy activities can be designed to foster spirituality skills among primary school children.

### **Acknowledgments**

There is no funding body which supported the present stud and there is no conflict of interest between the authors. Furthermore, all necessary, official and ethical permissions were taken from the local education authorities before the present study was launched.

The present study is a part of doctoral dissertation of Impact of Emotional Literacy Training on Student's Emotional Intelligence Level in Primary Schools by Kerem Coskun conducted at Ondokuz Mayıs University Graduate School of Education, under supervision of Yücel Oksuz.

### **ORCID**

Kerem Coskun  <https://orcid.org/0000-0002-3343-2112>

Yücel Öksüz  <https://orcid.org/0000-0001-9310-7506>

### **5. REFERENCES**

- Arda, T. B., & Ocak, S. (2012). Social Competence and Promoting Alternative Thinking Strategies PATHS Preschool Curriculum. *Educational Sciences: Theory and Practice*, 12(4), 2691-2698.
- Boyatzis, R. E., Goleman, D., & Rhee, K. (2000). Clustering competence in emotional intelligence: Insights from the Emotional Competence Inventory (ECI). In R. Bar-on, J.D.A. Parker (Eds.), *Handbook of emotional intelligence* (pp: 343-362). San Francisco: Jossey-Bass.
- Bredacs, A. (2010). emotional intelligence and its development in school-with a special view to talent management. *Practice and Theory in Systems of Education*, 5, 65-86.
- Brown, R. B. (2003). Emotions and behavior: Exercises in emotional intelligence. *Journal of Management Education*, 27(1), 122-134.
- Burman, E. (2009). Beyond emotional literacy in feminist and educational research. *British Educational Research Journal*, 35(1), 137-155.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 122, 155–159.
- Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education*. Oxon: Taylor & Francis Group.

- Coppock, V. (2007). It's good to talk! A multidimensional qualitative study of the effectiveness of emotional literacy work in schools. *Children & Society*, 21(6), 405-419.
- Coskun, K., Oksuz, Y. & Yilmaz, H.B. (2017). Ten years emotional intelligence scale (TYEIS): Its development, validity, and reliability. *International Journal of Assessment Tools in Education*, 4(2), 122-133.
- Di Fabio, A., & Kenny, M.E. (2011). Emotional intelligence and perceived social support among Italian high school students. *Journal of Career Development*, 39(5), 461-475.
- Dolev, N. (2012). *Emotional Intelligence Competencies in Teachers through Group-Based Coaching*. (Unpublished Doctoral Thesis). The University of Leicester, Leicester.
- Domitrovich, C.E., Cortes, R.C. & Greenberg, M.T. (2007). Improving young children's social and emotional competence: A randomized trial of the preschool paths curriculum. *The Journal of Primary Prevention*, 28(2), 67-91.
- Dulewicz, V., Higgs, M. (2004). Can emotional intelligence be developed? *The International Journal of Human Resource Management*, 15(1), 95-111.
- Emmons, R. A. (2000). Is spirituality an intelligence? Motivation, cognition, and the psychology of ultimate concern. *The International Journal for The Psychology of Religion*, 10(1), 3-26.
- Faupel, A. (2003). *Emotional literacy assessment and intervention ages 11-16*. Southampton: Nfer-Nelson Publishing.
- Ferrando, M., Prieto, M.D., Almeida, L.S., Ferrandiz, C., Bermejo, R., Lopez-Pina, J.A., Hernandez, D., Sainz, M., & Fernandez, C. (2010). Trait emotional intelligence and academic performance: Controlling for the effects of IQ, personality, and self-concept. *Journal of Psychoeducational Assessment*, 20(10), 1-10.
- Field, A. (2009). *Discovering Statistics Using SPSS*. London: Sage Publications.
- Flynn, S.J. (2010). *A comparative and exploratory study of the nfer-nelson emotional literacy scale in an Irish context* (Unpublished Doctoral Dissertation). The University of Exeter, Exeter.
- Fraenkel, F., Wallen, N.E., & Hyun, H. (2012). *How to design and evaluate research in education*. New York: McGraw Hill.
- Gillum, J. (2010). *Using emotional literacy to facilitate organizational change in a primary school: A case study* (Unpublished Doctoral Dissertation). The University of Birmingham, Birmingham.
- Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ*. New York: Bantam Books.
- Goleman, D. (1998). *Working with emotional intelligence*. New York: Bantam Books.
- Greenberg, M.T., & Kushe, C.A. (1998). Preventive intervention of school-age deaf children: The paths curriculum. *Journal of Deaf Studies and Deaf Education*, 3(1), 49-63.
- Haddon A, Goodman H, Park J, & Crick R.D. (2005). Evaluating Emotional Literacy in Schools: The Development of the School Emotional Environment for Learning Survey. *Pastoral Care in Education*, 23(4), 5-16.
- Hallam, S. (2009). An Evaluation of the social and emotional aspects of learning (SEAL) program: Promoting positive behaviour, effective learning and well-being in primary school children. *Oxford Review of Education*, 35(3), 313-330.
- Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2012). Promoting young children's social competence through the preschool PATHS curriculum and MyTeachingPartner professional development resources. *Early Education & Development*, 23(6), 809-832.
- Jacobson, M., & Ruddy, M. (2004). *Open to outcome: A practical guide for facilitating and teaching experiential reflection*. Oklahoma: Wood'N'Barnes Publishing.

- Jellesma, F.C., Rieffe, C., Terwogt, M.M., Westenberg, M. (2011). Children's sense of coherence and trait emotional intelligence: A longitudinal study exploring the development of somatic complaints. *Psychology & Health, 26*(3), 307-320.
- Jones, S.M., Brown, J.L., & Aber, L.J. (2011). Two-year impacts of universal school-based social-emotional and literacy intervention: An experiment in translational and developmental research. *Child Development, 82*(2), 533-554.
- Keefer, K.V., Holden, R.R., & Parker, J.D.A. (2013). Longitudinal assessment of trait emotional intelligence: measurement invariance and construct continuity from late childhood to adolescence. *Psychological Assessment, 25*(4), 1255-1272.
- Kelly, B., Longbottom, J., Potts, P., Williamson, J. (2004). Applying Emotional Intelligence: Exploring the Promoting Alternative Thinking Strategies Curriculum. *Educational Psychology, 20*(3), 221-240.
- Kolb, D.A. (1984). *Experiential learning: Experience as the source of learning and development*. New Jersey: Prentice Hall.
- Lu, C., Buchanan, A. (2014). Developing students' emotional well-being in physical education. *Journal of Physical Education, Recreation & Dance, 85*(4), 28-33.
- Matthews, B. (2006). *Engaging education: Developing emotional literacy, equity and co-education*. Berkshire: Open University Press.
- O'Hara, D. (2011). The impact of peer mentoring on pupils' emotional literacy competencies. *Educational Psychology in Practice, 27*(3), 271-291.
- Orbach, S. (1998). Emotional literacy. *Young Minds Magazine, 33*, 12-13.
- Paek, E. (2006). Religiosity and perceived emotional intelligence among Christians. *Personality and individual differences, 41*(3), 479-490.
- Park, J. (1999). Emotional literacy: Education for meaning. *International Journal of Children's Spirituality, 4*(1), 19-28.
- Perry, L., Lennie, C., & Humphrey, N. (2008). Emotional literacy in the primary classroom: teacher perceptions and practices. *Education 3-13, 36*(1), 27-37.
- Pratt, L. (2009). *Supporting pupils at risk of exclusion: an evaluation of an intensive, out of school, emotional literacy program for key stage 3 pupils* (Unpublished Doctoral Thesis). Institute of Education, University of London, London.
- Qualter, P., Whiteley, Y., Hutchinson, J.M., & Pope, D.J. (2007). Supporting the development of emotional intelligence competencies to ease the transition from primary to high school. *Educational Psychology in Practice, 23*(1), 79-95.
- Reddy, L.A., & Richardson, L. (2006). School based prevention and intervention programs for children with emotional disturbance. *Education and Treatment of Children, 29*(2), 379-404.
- Rivers, S.E., Brackett, M.A., Reyes, M.R., Mayer, J.D., Caruso, D.R., & Salovey, P. (2012). Measuring emotional intelligence in early adolescence with the MSCEIT-YV psychometric properties and relationship with academic performance and psychosocial functioning. *Journal of Psychoeducational Assessment, 30*(4), 344-366.
- Saltalı, N. D. (2010). *Duygu eğitiminin okul öncesi dönem çocuklarının duygusal becerilerine etkisi* (Unpublished Doctoral dissertation) Selçuk University, Institute of Social Sciences. Retrieved from <http://acikerisim.selcuk.edu.tr:8080/xmlui/bitstream/handle/123456789/7044/258519.pdf?sequence=1&isAllowed=y>
- Shapiro, S.S. & Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3), 591-611.
- Skinner, B.F. (1984). Selection by consequences. *Behavioural and Brain Sciences, 7*, 477-481.
- Steiner, C. (1979). *Emotional literacy: Intelligence with a heart*. California: Personhood Press.

- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Allyn & Bacon/Pearson Education.
- Tew, M. (2007). *School effectiveness: Supporting success through emotional literacy*. London: Sage Publications.
- Tufan, Ş. (2011). *Geliştirilen duygusal zekâ eğitim programının ortaöğretim dokuzuncu sınıf öğrencilerinin duygusal zekâ düzeylerine etkisi*. (Unpublished master thesis). Ankara University, Institute of Educational Sciences.
- Yaşarsoy, E. (2006). *Duygusal zeka gelişim programının, eğitilebilir zihinsel engelli öğrencilerin davranış problemleri üzerindeki etkisinin incelenmesi*. (Unpublished master thesis). Çukurova University, Institute of Social Sciences.
- Zeidner, M., Roberts, R.D., & Matthews, G. (2002). Can emotional intelligence be schooled? A critical review. *Educational Psychologist*, 37(4), 215-231.



## An Investigation of Item Bias of English Test: The Case of 2016 Year Undergraduate Placement Exam in Turkey

Rabia Akcan <sup>1,\*</sup>, Kübra Atalay Kabasakal <sup>2</sup>

<sup>1</sup> Ministry of National Education, 03700, Afyonkarahisar, Turkey

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

### ARTICLE HISTORY

Received: 25 September 2018

Revised: 20 December 2018

Accepted: 02 January 2019

### KEYWORDS

Undergraduate Placement Exam,  
differential item functioning,  
differential bundle functioning,  
item bias,  
MIMIC

**Abstract:** The purpose of this study is to determine whether English test items of Undergraduate Placement Exam (UPE) in 2016 contain differential item functioning (DIF) and differential bundle functioning (DBF) in terms of gender and school type and examine the possible sources of bias of DIF items. Mantel Haenszel (MH), Simultaneous Item Bias Test (SIBTEST) and Multiple Indicator and Multiple Causes (MIMIC) methods were used for DIF analyses. DBF analyses were conducted by MIMIC and SIBTEST methods. Expert opinions were consulted to determine the sources of bias. Data set of the study consisted of responses of 59818 students to 2016 UPE English test. As a result of the analyses carried out on 60 items, it was seen that one item in translation subtest contained DIF favoring male students. In school type based analyses, it was concluded that there were nine DIF items in vocabulary and grammar knowledge subtest, six DIF items in reading comprehension subtest and four DIF items in translation subtest. Experts stated that one item containing DIF by gender was unbiased, and evidence of bias was found in thirteen of nineteen items that contained DIF by school type. According to DBF analyses, it was found that some item bundles contained DBF with respect to gender and school type. As a result of research, it was discovered that there were differences with regard to the number of DIF items identified by three methods and the level of DIF that the items contained; however, methods were consistent in detecting uniform DIF.

## 1. INTRODUCTION

Large scale tests are commonly used throughout the world with the aim of selection and placement of the students. To make fair and right decisions based on the test results, and select students who have the ability and interest in accordance with the departments, the ability to be measured in the test must be evaluated accurately. Hence, it is significant to have well-qualified items for the tests. In a test, probability of answering an item correctly must not be influenced

CONTACT: Rabia Akcan  [eltrabia42@hotmail.com](mailto:eltrabia42@hotmail.com)  Ministry of National Education, 03700, Çay/Afyonkarahisar, Turkey

ISSN-e: 2148-7456 / © IJATE 2019



by variables such as examinees' socio economic status, gender or school type they studied. Otherwise, test becomes biased and might not reflect examinees' cognitive abilities.

Bias is described as systematic errors in measurement process (Osterlind, 1983). The concept of bias is directly associated with fairness, and it is the condition in which group characteristics that are not related to the construct to be measured affect test results. Thus, test bias distorts results by allowing examinees' characteristics influence measurement of main construct. Consequently, test measures an irrelevant construct in addition to the intended one (Mcnamara & Roever, 2006). When a group of examinees has a higher probability of responding an item correctly than another group due to some characteristics of the item or inconvenient test conditions, it is called item bias (Zumbo, 1999). Item bias is a possible threat to validity (Clauser & Mazor, 1998). Therefore, doing research on this matter is of importance.

Language teaching has become increasingly important throughout the world. Countries develop language tests for measuring language skills of the students, and decisions are made based on test results. Because these tests shape the future of the students as well as the countries, preparing equal and valid tests is highly significant. Since English is an international language and frequently used in science and technology, in this study English test in 2016 Undergraduate Placement Exam (UPE) held in Turkey was examined in terms of item bias. A test item can be said to be biased when it is in favor of one group and to the disadvantage of another group. These items show differential item functioning (DIF). DIF occurs when testtakers from different groups have different probability of success on an item after they are matched on the ability to be measured. DIF is a necessary condition but containing DIF is not sufficient for item bias (Clauser & Mazor, 1998). DIF can be present in two forms as uniform and non-uniform. When a group of examinees has higher likelihood of answering an item correctly than another group across all ability levels, uniform DIF occurs (Finch, 2005). On the other hand, non uniform DIF is present if the difference of the likelihood of answering an item correctly between the two groups is inconsistent across all ability levels (Camilli & Shepard, 1994).

Although the focus is generally on the single item DIF analysis, there are many tests consisting of small item bundles. An item bundle is described as a set of items selected according to an organizing principle and these items do not have to be adjacent and they are not necessarily related to a text or passage. When DIF analysis is conducted on item bundles, it is called differential bundle functioning (DBF) (Douglas, Roussos, & Stout, 1996). In the literature there are some studies on DIF in language tests however; more research is needed to improve test quality.

Lin and Wu (2003) investigated DIF and DBF with respect to gender in English Proficiency Test used in China. Simultaneous Item Bias Test (SIBTEST) method was used for the analyses. Research results revealed that two items contained large DIF, and eleven items contained moderate DIF. Four of these items were listening items favoring females and three of them were grammar and vocabulary items favoring males. Two cloze test items and three reading items also favored males, and only one reading item was in favor of females. According to DBF results, one listening item bundle favored females systematically, and the other item bundles favored males slightly.

Abbott (2007) carried out DIF and DBF analyses of reading passages separated according to bottom-up and top-down strategies. SIBTEST method was used for the analyses. Hypothesis of the research was based on the claim that Chinese students are more successful in bottom-up strategies and Arabic students are more successful in top-down strategies. In analyses, items were separated into two categories in line with these two strategies. Research results showed that there were significant systematic differences between the two groups in using these strategies.

Kan (2007) conducted DIF analysis of items used in Hacettepe University foreign language proficiency examination. DIF analyses were carried out in terms of gender and the departments by using Mantel Haenszel (MH) method. It was reported that one item showed DIF favoring female students. Twelve items contained DIF in terms of departments separated into three categories as social sciences, physical sciences and health sciences.

Karakaya and Kutlu (2012) investigated item bias of Turkish subtests in Level Determination Exam. DIF analyses were conducted in terms of gender and school type using MH and Logistic Regression methods. Expert opinions revealed that only one item (item 19) in 8th grade Turkish subtest was biased in favor of males. Experts stated that item 19 included some expressions associated with feeding fish in an aquarium. Since male students are more interested in aquariums and feeding fish, item 19 was identified as biased.

Although there are many DIF detection methods described in literature, very few of them are used in practice (Clauser & Mazor, 1998). These methods can be broadly categorized into two as Classical Test Theory (CTT) and Item Response Theory (IRT) based methods. Camilli and Shepard (1994) stated that Confirmatory Factor Analysis (CFA) methods can also be used in DIF detection. In this study, MH, SIBTEST and Multiple Indicators Multiple Causes (MIMIC) methods were used for DIF detection.

### 1.1. Mantel-Haenszel

Mantel Haenszel (MH) statistic was proposed by Holland and Thayer (1988) and it has been commonly used for DIF detection since then. In this method, two groups are matched on the ability, and the probability of success on the item is compared between groups. Total test scores are generally used for matching (Clauser & Mazor, 1998). Afterwards, reference and focal groups are matched on total test scores, and a 2x2xS contingency table is created. S represents the different number of total test score. At all ability levels, data for each item can be organized as in Table 1 (Roussos & Stout, 1996).

**Table 1.** MH Method Data Organization.

| Group     | 1 =Correct | 0=Incorrect | Total    |
|-----------|------------|-------------|----------|
| Reference | $A_j$      | $B_j$       | $N_{Rj}$ |
| Focal     | $C_j$      | $D_j$       | $N_{Oj}$ |
| Total     | $M_{1j}$   | $M_{0j}$    | $T_j$    |

Using these tables that are formed at all ability levels, likelihood ratio ( $\alpha$ ) is estimated and this ratio is shown by equation 1 (Clauser & Mazor, 1998).

$$\alpha = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j} \quad (1)$$

To facilitate interpretation, log of  $\alpha$  is taken and resulting value is multiplied by -2.35. Thus  $\Delta_{MH}$  is produced. Positive values of  $\Delta_{MH}$  show DIF against reference group and negative values of  $\Delta_{MH}$  show DIF against the focal group (Clauser & Mazor, 1998). Zieky (1993) classified  $\Delta_{MH}$  statistic as the following:  $|\Delta_{MH}| < 1$  indicates negligible DIF,  $1 \leq |\Delta_{MH}| \leq 1.5$  indicates moderate DIF and  $|\Delta_{MH}| \geq 1.5$  indicates large DIF.

### 1.2. SIBTEST

SIBTEST was developed by Shealy and Stout (1993) and is based on the standardization procedure. In this method, test items are separated into two groups as studied subtest and matching subtest. Corresponding matching subtest scores for the reference and focal groups are

estimated for each matching subtest, and these scores are modified using regression correction. Lastly, the ratio of answering the studied item correctly for reference and focal group is estimated. By using the weighed sum of the difference between these ratios,  $\beta$  parameter is found (Roussos & Stout, 1996).

SIBTEST hypothesis is given by:

$$H_0: \beta = 0 \quad H_1: \beta \neq 0 \quad (2)$$

And the size of DIF is expressed as:

$$\beta = \int [P(\theta, R) - P(\theta, F)] f_F(\theta) d\theta \quad (3)$$

where  $P(\theta, R)$ , probability of correct response for examinees from reference group;  $P(\theta, F)$ , probability of correct response for examinees from focal group;  $f_F(\theta)$ , density function in focal group;  $d$  is the width of the scaling interval. With SIBTEST method, items or item bundles in the secondary dimension can be detected, and DIF analysis can be carried out.  $\beta$  parameter is used to identify the size of DIF for items or item bundles (Gierl & Khaliq, 2001).

Roussos and Stout (1996) proposed guidelines for  $\beta$  parameter to classify the size of DIF that have three levels : negligible DIF ( $|\beta| < 0,059$ ), moderate DIF ( $0,059 \leq |\beta| \leq 0,088$ ) and large DIF ( $|\beta| > 0,088$ ). Positive values of  $\beta$  show DIF against focal group and negative values of  $\beta$  show DIF against the reference group. However, no guidelines have been proposed for classifying the size of DBF (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001).

In this research, both uniform and nonuniform DIF have been identified using SIBTEST method. Li and Stout (1996) proposed Crossing-SIBTEST (CSIBTEST) statistic which they see as a better alternative to SIBTEST statistic for identifying nonuniform DIF. This statistic was modified by Chalmers (2017), and it was stated that modified version of CSIBTEST statistic can be used in place of the original CSIBTEST statistic. While data can be analysed with samples consisted of at most 7000 individuals for reference and focal groups in SIBTEST programme, there is no limit in R software. Therefore; DIF analysis was performed with “mirt” package (Chalmers, 2018) in R which gives an opportunity to estimate SIBTEST and CSIBTEST statistics simultaneously.

### 1.3. MIMIC

MIMIC is a model of CFA and can be used to detect DIF. MIMIC models estimate direct and indirect effects for a grouping variable. Latent trait is regressed onto grouping variable by indirect effect to show whether there is group mean differences on the latent trait. By direct effect, item responses are regressed onto grouping variable to find out whether response probabilities differ across groups (Finch, 2005).

MIMIC model in DIF context is expressed as:

$$y^*i = \lambda_i \eta + \beta_i z_k + \varepsilon_i \quad (4)$$

where  $y^*i$ , latent response variable;  $\lambda_i$ , factor loading for variable  $i$ ;  $\eta$ , latent trait;  $\beta_i$ , slope relating the group variable with the response;  $\varepsilon_i$ , random error;  $z_k$ , a dummy variable showing group membership (Finch, 2005).

Finch (2012) expanded MIMIC model, and used MIMIC as an alternative to SIBTEST in identifying DBF. Results of the research revealed that MIMIC model was as effective as SIBTEST in detecting DBF. In analyses with MIMIC method, positive values of beta show DIF against the group coded as 0, negative values of beta show DIF against the group coded as 1.

In this study focal group was coded as 0, and reference group was coded as 1. No criterion has been proposed to determine the size of DIF and DBF with MIMIC method.

## 2. METHOD

This study is a descriptive research as it investigates DIF and DBF of English test items in UPE, and it is also a qualitative research because it examines the possible sources of bias in DIF items. English test includes 80 items; however, 20 testlet items were excluded from the analysis. Therefore; 60 dichotomous scored items were analysed in terms of gender and school type.

### 2.1. Population and Sample

Population of the study consists of 88284 examinees who took English test in 2016 year UPE. Data set included 87 school types, and 74 of them were not analysed because a part of them was less than 1% of data set and the others were shut down by Ministry of National Education. Rest of the schools was gathered under four school types as they have similar educational objectives. Those four schools are vocational high school (VHS), Anatolian high school (AHS), religious vocational high school (RVHS) and private high school (PHS). Before factor analysis is conducted, data set should be checked whether it is appropriate for the analysis. To accomplish this, it was determined whether the data set included missing values and outliers. It was seen that missing values were below 5% of the data set, and zero imputation was used for the missing values. Data set was also examined in terms of univariate and multivariate outliers, and it was found that there were 1853 multivariate outliers in the data. Analyses were carried out using data from 59818 examinees after these outliers were removed. Distribution of data according to gender and school type is reported in [Table 2](#).

**Table 2.** Gender and School Type Distribution.

| Group        | N     | %    |
|--------------|-------|------|
| Gender       |       |      |
| Female       | 36101 | 60.4 |
| Male         | 23717 | 39.6 |
| School Type  |       |      |
| Vocational   | 10140 | 17.0 |
| Anatolian    | 21618 | 36.1 |
| Religious V. | 11194 | 18.7 |
| Private      | 16866 | 28.2 |
| Total        | 59818 | 100  |

### 2.2. Instrument

English test in UPE consists of three parts as vocabulary and grammar knowledge (15 items), reading comprehension (48 items) and translation (12 items). All items in the test are multiple choice items. Vocabulary and grammar knowledge part includes items that measure basic vocabulary and grammar knowledge of the students. Reading comprehension part contains seven different item types. These are paragraph completion, cloze test, reading paragraphs, dialogue completion, sentence completion, irrelevant sentence and situational dialogue. Translation part consists of English-Turkish translation items and Turkish-English translation items. Exam takes two hours. Students are placed in departments according to their results.

### 2.3. Data Set

Data set used in this research was obtained from Research and Development Unit of Student Selection and Placement Center.

## 2.4. Data Analysis

Test items on the data set were scored as 1 for correct response and 0 for wrong or blank response. To examine the structure of the data, factor analysis was made using “lavaan” package (Rosseel, 2017) in R. It was made based on tetrachoric correlation matrix, and parallel analysis was used to decide the number of factors. To accomplish this “polycor” package (Fox, 2016) and “nFactors” package (Raiche & Magis, 2015) were used. Factor analysis indicated that there were 3 dimensions which were vocabulary and grammar knowledge (items 1-15), reading comprehension (items 21-28, 44-63 and 76-80) and translation (items 64-75) dimensions. Descriptive statistics and item statistics according to groups and DIF analyses were carried out based on these dimensions. DIF analyses were performed by using SIBTEST, MH and MIMIC methods. Descriptive statistics and item statistics were estimated using “CTT” package (Willse, 2018), and DIF analyses with SIBTEST were conducted using “mirt” package (Chalmers, 2018), MH analyses were performed using “difR” package (Magis, Beland, & Raiche, 2016) in R. DIF analyses with MIMIC were carried out using Mplus (Muthen & Muthen, 1998) via “MplusAutomation” package (Hallquist & Wiley, 2018) in R. To determine the possible sources of bias in DIF items expert opinions were consulted. SIBTEST and MIMIC methods were used for DBF analyses as well.

## 3. FINDINGS

### 3.1. DIF Results and Expert Opinions

In this study, DIF and DBF analyses of English test of UPE in 2016 was conducted in terms of gender and school type. Analyses were carried out based on three dimensions regarded as subtests. Six comparisons were made in each subtest with regard to school type. Items that show moderate or large DIF with SIBTEST and MH methods were considered as DIF items if they show DIF with MIMIC method at the same time. Ten experts were consulted to determine the possible sources of bias. Four experts work as English language teachers in the Ministry of National Education. Two experts have a degree of doctoral philosophy in department of English teaching, and four experts have a degree of doctoral philosophy in educational measurement and evaluation. DIF analyses results according to gender are reported in [Table 3](#).

**Table 3.** DIF Items by Gender in Each Subtest

| Subtests/Gender                  | Female | Male |
|----------------------------------|--------|------|
| Vocabulary and grammar knowledge | -      | -    |
| Reading comprehension            | -      | -    |
| Translation                      | -      | 68   |

\*DIF items that contain nonuniform DIF with SIBTEST method

As shown in [Table 3](#), there are no common DIF items with three methods in vocabulary and grammar knowledge and reading comprehension subtests. There is one DIF item in favor of males in translation subtest. Experts stated that there is no evidence of bias in item 68 favoring males. Results of DIF analyses with regard to school type are given in [Table 4](#).

[Table 4](#) shows that there are no common items with three methods in reading comprehension and translation subtests in VHS-AHS comparison. However, in vocabulary and grammar knowledge subtest items 3, 5 and 8 indicated DIF in favor of VHS. It was found that items 7 and 12 contained DIF in favor of AHS with MH and MIMIC methods. These items showed nonuniform DIF with SIBTEST method. Five experts pointed out that as item 3 included some expressions related to information technology this may have given an advantage to the students graduated from information technology (IT) departments of VHS. Because those students are familiar with the expressions and this could be a possible source of bias in the item. Whereas



seven experts stated that there is no evidence of bias in items 5 and 8, three experts identified different sources of bias such as materials used in classes, familiarity with the expressions used in items and knowing problem solving techniques that can help students answer the items easily.

**Table 4.** DIF Items by School Type in Each Subtest

|                                  |                |                |
|----------------------------------|----------------|----------------|
| Subtests/School Type             | Vocational     | Anatolian      |
| Vocabulary and grammar knowledge | 3, 5, 8        | 7*,12*         |
| Reading comprehension            | -              | -              |
| Translation                      | -              | -              |
| Subtests/ School Type            | Vocational     | Religious Voc. |
| Vocabulary and grammar knowledge | -              | 13             |
| Reading comprehension            | 57, 58         | -              |
| Translation                      | -              | -              |
| Subtests/ School Type            | Vocational     | Private        |
| Vocabulary and grammar knowledge | 3, 8, 13*      | 12*            |
| Reading comprehension            | -              | -              |
| Translation                      | 71*            | -              |
| Subtests/ School Type            | Anatolian      | Religious Voc. |
| Vocabulary and grammar knowledge | 7*,10*,12*,14* | 3,4*,5,8,13    |
| Reading comprehension            | 57, 58         | -              |
| Translation                      | -              | -              |
| Subtests/ School Type            | Anatolian      | Private        |
| Vocabulary and grammar knowledge | 7*             | -              |
| Reading comprehension            | -              | 26, 47         |
| Translation                      | 72             | 68             |
| Subtests/ School Type            | Religious Voc. | Private        |
| Vocabulary and grammar knowledge | 3, 8, 13       | 10*, 12*       |
| Reading comprehension            | 63             | 45*, 57, 58    |
| Translation                      | 72*            | 73*            |

\*DIF items that contain nonuniform DIF with SIBTEST method

When Table 4 was examined, it was seen that AHS students were advantageous in items 7 and 12 in comparisons with VHS and RVHS. For item 7, six experts asserted that students graduated from AHS may have been frequently exposed to that type of items and they might be familiar with grammatical structure used in that item. This could be the reason for the difference between AHS and the other school types. For item 12, one of the experts stated that science-related terms used in the item might have helped AHS students understand the item easily. On the other hand, five experts agreed on the idea that AHS students are familiar with the grammatical structures like “unless”, used in item 12, and thus this may have given them an advantage. Four experts found no evidence of bias in item 12.

In VHS- RVHS comparison, item 13 in vocabulary and grammar knowledge subtest contained DIF favoring RVHS. Items 57 and 58 in reading comprehension subtest showed DIF in favor of VHS. However, there is no common item with three methods in translation subtest. According to experts, item 13 is unbiased. Nine experts found no evidence of bias in item 57 and one expert claimed that students graduated from cookery department of VHS may have been familiar with the terms used in the item, and that might be a source of bias. Three experts stated that item 58 is situational dialogue item, and RVHS students might be unfamiliar with the situation given in the item due to socioeconomic and cultural differences. No evidence of bias was identified in item 58 by the other seven experts.



In VHS-PHS comparison, items 3, 8 and 13 in vocabulary and grammar knowledge subtest showed DIF favoring VHS and item 12 showed DIF favoring PHS. While there is no common item containing DIF with three methods in reading comprehension subtest, one item (71) contained DIF favoring VHS in translation subtest. It was determined that items 13 and 71 showed DIF in favor of VHS and item 12 contained DIF in favor of PHS with MH and MIMIC methods. These items showed nonuniform DIF with SIBTEST method.

As mentioned before five experts stated that item 3 may have favored VHS students because it includes some expressions that IT students can understand more easily. Item 8 was found unbiased by seven experts. Yet two experts asserted that VHS students may have answered this item by just choosing simple option. For item 12, two experts pointed out those science-related words like “body cells” might be the source of bias and three experts found no source of bias. Five experts explained that PHS students might have been exposed to and be familiar with that type of items as in AHS students’ case, so this could be the reason of the difference between VHS-PHS. All experts had a common view that items 13 and 71 were not biased.

In AHS-RVHS comparison, four items were in favor of AHS and five items were in favor of RVHS in vocabulary and grammar knowledge subtest. Two items favored AHS in reading comprehension subtest, and there were no common items showing DIF with three methods in translation subtest. While items 7, 10, 12 and 14 had DIF favoring AHS and item 4 showed DIF favoring RVHS with MH and MIMIC methods, they contained nonuniform DIF with SIBTEST method.

As in items 7 and 12 mentioned in AHS-VHS comparison, six experts stated that item 14 is a grammar item and it includes the frequently used expression “not only, but also”, thus AHS students may have practiced, and they became familiar with that type of items which might be a factor that made AHS students more successful on the item. As for item 10, four experts explained that AHS students are more interested in science, so some expressions used in item 10 such as “brain”, “scientific evidence” and science related content of the item might have given an advantage to AHS students. Whereas three experts considered familiarity with item type as a source of bias, three experts found no source of bias in the item.

Six experts stated that item 57 was not biased. On the other hand, four experts said that item 57 is a situational dialogue item and RVHS students may not be familiar with the sample situation in the item as it includes the words “vegetarian”, “beefsteak” etc. Socioeconomic and cultural differences were proposed as the cause of bias. Similarly, item 58 was determined to be biased by three experts as it contains a situational dialogue that is not familiar with RHVS students. For item 3, only one expert asserted that item might be biased because it requires students to remember the information and RVHS students are mostly educated based on rote learning. Item 5 was found to be unbiased by seven experts, and one expert stated that this item also requires memory like item 3. Socioeconomic differences were defined as the cause of bias by two experts. For item 8, four experts pointed out that words used in the item such as “temple” and “dome” might create the difference between the schools because religious terms may have made the item easier to understand for RHVS students. Two experts explained that they might have responded the item just choosing the simple distractor as well. No evidence of bias was found by experts in items 4 and 13 between AHS-RVHS comparison.

In AHS-PHS comparison, item 7 showed DIF favoring AHS in vocabulary and grammar knowledge subtest and items 26 and 47 contained DIF favoring PHS in reading comprehension subtest. Items 68 and 72 indicated DIF favoring PHS and AHS, respectively. Eight experts found no evidence of bias in item 26; however, one expert emphasized the importance of language teaching techniques in PHS. Because language is taught for everyday use and students get a chance to practice it, PHS students might be advantageous in this sentence completion item. Moreover, one expert stated that item contains expressions related to science such as

“muscular pump”, “blood flooding” and heart”, so this could be a reason for the difference between schools. For item 47, a dialogue completion item, four experts explained that language education based on practice and everyday use in PHS may have given an advantage to PHS. Scientific content of the item and eating habits were proposed as sources of bias in that item by two experts. Item 72 was found to be unbiased by experts. Six experts found no evidence of bias in item 68 and four experts pointed out that students in PHS with a higher socioeconomic level might get the item more easily due to their social life and family structure because it includes the words “French and British antique”, “antiques bazaar” and “antiques lovers”.

In RVHS-PHS comparison, items 3, 8 and 13 showed DIF favoring RVHS and items 10 and 12 showed DIF favoring PHS in vocabulary and grammar knowledge subtest. Items 45, 57 and 58 contained DIF favoring PHS in reading comprehension subtest. Item 72 had DIF favoring RVHS and item 73 had DIF favoring PHS. But, items 10, 12, 45, 72 and 73 indicated nonuniform DIF with SIBTEST method.

When Table 4 was examined, it was seen that items 3, 8 and 13 were in favor of RVHS in RVHS-PHS comparison as they were in AHS-RVHS comparison. Therefore, experts stated that the sources of bias mentioned earlier in AHS-RVHS comparison are also valid in RVHS-PHS comparison. Likewise, AHS and PHS students were more advantageous in items 57 and 58 than RVHS students. Hence, for these items experts showed the same sources of bias given in AHS-RVHS comparison. For item 10, five experts showed the scientific content of the item and three experts showed the familiarity with the item type as a source of bias because PHS students are more likely to encounter that type of items and they tend to learn science. Similarly, type of the item and scientific expressions used in item 12 were indicated as a source of bias by experts. It was clearly seen that AHS and PHS students were advantageous in item 12.

For item 45, seven experts pointed out that expressions used in the item such as “technology”, “futureFest”, “demos” and “innovation” might have given an advantage to PHS students because they are more likely to attend festivals and have an idea about them thanks to their socioeconomic level. Two experts stated that practice and educational activities performed in PHS might affect the results as item 45 is a dialogue completion item. Item 73 was considered as unbiased by six experts. Two experts explained that activities about different countries and cultures may have been done more in PHS than RVHS, and the other two experts showed practice and education based on every day use as the cause of difference between RVHS-PHS. Experts reached a consensus on that items 63 and 72 did not have bias.

### 3.2. DBF Results

Item bundles to be analysed can be chosen using different methods such as content analysis, table of specifications or psychological analysis (Gierl, Bisanz, Bisanz, Boughton & Khaliq, 2001). Nevertheless, UPE English test consists of item bundles including different types of items. These item bundles are prepared to measure instructional objectives and cognitive abilities of the students. There are two item bundles in vocabulary and grammar knowledge and translation subtests and six item bundles in reading comprehension subtest. Item bundles and item numbers are given in Table 5.

DBF analyses were carried out using SIBTEST and MIMIC methods. Woods and Grimms (2011) reported that MIMIC model was used to detect nonuniform DIF in their research and it worked better than the other model but type I error of this model was highly inflated. Although MIMIC can be used to identify nonuniform DIF or DBF, in this research only uniform DIF and DBF were detected by MIMIC owing to inflated type I error.

**Table 5.** Item Bundles and Numbers of Items

| Item Bundles                     | Item Numbers                       |
|----------------------------------|------------------------------------|
| Vocabulary and Grammar Knowledge |                                    |
| Vocabulary                       | 1, 2, 3, 4, 5                      |
| Grammar                          | 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 |
| Reading Comprehension            |                                    |
| Sentence Completion              | 21, 22, 23, 24, 25, 26, 27, 28     |
| Dialogue Completion              | 44, 45, 46, 47, 48                 |
| Paraphrasing                     | 49, 50, 51, 52, 53                 |
| Situational Dialogue             | 54, 55, 56, 57, 58                 |
| Paragraph Completion             | 59, 60, 61, 62, 63                 |
| Irrelevant Sentence              | 76, 77, 78, 79, 80                 |
| Translation                      |                                    |
| English-Turkish Translation      | 64, 65, 66, 67, 68, 69             |
| Turkish- English Translation     | 70, 71, 72, 73, 74, 75             |

Furthermore, whether matching subtest contains DIF or not is an important issue in DBF analysis. Finch (2005) stated that DIF items found in matching subtest may threaten the accuracy of statistical methods in identifying DBF. Results of DIF analyses by school type demonstrated that there were quite a number of items including moderate or large DIF in vocabulary and grammar knowledge and translation subtests. Hence, DBF analyses by school type were only performed in reading comprehension subtest. DBF results by gender are given in [Table 6](#).

**Table 6.** DBF Results by Gender

| Methods/Item Bundles         | SIBTEST | MIMIC |
|------------------------------|---------|-------|
| Vocabulary                   | M       | M     |
| Grammar                      | F       | -     |
| Sentence Completion          | M       | M     |
| Dialogue Completion          | NU      | -     |
| Paraphrasing                 | NU      | -     |
| Situational Dialogue         | F       | F     |
| Paragraph Completion         | M       | M     |
| Irrelevant Sentence          | F       | F     |
| English-Turkish Translation  | M       | F     |
| Turkish- English Translation | F       | F     |

\*F: Female, M: Male, NU: Non-uniform DIF

As shown in [Table 6](#), seven item bundles showed DBF with both methods. However, methods are inconsistent in English-Turkish translation bundle. DBF results by school type are given in [Table 7](#).

When [Table 7](#) was examined, it was found that all bundles in AHS-VHS and RVHS-PHS comparisons indicated DBF with both methods. Five item bundles showed DBF in VHS-RVHS and AHS-RVHS comparisons with both methods. In VHS-RVHS comparison three item bundles had DBF, and in AHS-PHS comparison there were four common item bundles showing DBF with both methods. Nevertheless, DBF results of two methods do not completely comply with each other.

**Table 7.** DBF Results by School Type

| Methods/Item Bundles (VHS-AHS)  | SIBTEST | MIMIC |
|---------------------------------|---------|-------|
| Sentence Completion             | NU      | AHS   |
| Dialogue Completion             | NU      | AHS   |
| Paraphrasing                    | VHS     | AHS   |
| Situational Dialogue            | AHS     | AHS   |
| Paragraph Completion            | NU      | AHS   |
| Irrelevant Sentence             | NU      | AHS   |
| Methods/Item Bundles (VHS-RVHS) | SIBTEST | MIMIC |
| Sentence Completion             | UN      | VHS   |
| Dialogue Completion             | UN      | -     |
| Paraphrasing                    | -       | -     |
| Situational Dialogue            | VHS     | VHS   |
| Paragraph Completion            | RVHS    | RVHS  |
| Irrelevant Sentence             | UN      | -     |
| Methods/Item Bundles (VHS-RVHS) | SIBTEST | MIMIC |
| Sentence Completion             | NU      | PHS   |
| Dialogue Completion             | NU      | PHS   |
| Paraphrasing                    | VHS     | -     |
| Situational Dialogue            | NU      | PHS   |
| Paragraph Completion            | VHS     | PHS   |
| Irrelevant Sentence             | NU      | PHS   |
| Methods/Item Bundles (AHS-RVHS) | SIBTEST | MIMIC |
| Sentence Completion             | NU      | AHS   |
| Dialogue Completion             | NU      | AHS   |
| Paraphrasing                    | RVHS    | -     |
| Situational Dialogue            | NU      | AHS   |
| Paragraph Completion            | RVHS    | AHS   |
| Irrelevant Sentence             | NU      | AHS   |
| Methods/Item Bundles (AHS-PHS)  | SIBTEST | MIMIC |
| Sentence Completion             | NU      | PHS   |
| Dialogue Completion             | PHS     | PHS   |
| Paraphrasing                    | AHS     | AHS   |
| Situational Dialogue            | NU      | AHS   |
| Paragraph Completion            | NU      | -     |
| Irrelevant Sentence             | AHS     | -     |
| Methods/Item Bundles (RVHS-PHS) | SIBTEST | MIMIC |
| Sentence Completion             | NU      | PHS   |
| Dialogue Completion             | NU      | PHS   |
| Paraphrasing                    | RVHS    | PHS   |
| Situational Dialogue            | PHS     | PHS   |
| Paragraph Completion            | RVHS    | PHS   |
| Irrelevant Sentence             | NU      | PHS   |

\* VHS: Vocational High School, AHS: Anatolian High School, RVHS: Religious Vocational High School, PHS: Private High School, NU: Non-uniform DIF

#### 4. DISCUSSION and CONCLUSION

In this study it was aimed to determine whether items of English test of UPE in 2016 show DIF and DBF in terms of gender and school type and examine the possible sources of bias of DIF items. MH, SIBTEST and MIMIC methods, which are based on CTT, IRT and CFA respectively, were used for analyses. It was reported in literature that detection methods are influenced by some factors such as sample size, proportion of DIF and ability difference among groups (Finch, 2005; Finch & French, 2007; Narayanan & Swaminathan, 1994). For this reason,

using different DIF detection methods increases reliability of research results. There are also some researches that suggest using more than one method to get more reliable results (Akin Arıkan, Uğurlu, & Atar, 2016; Gök, Kelecioğlu, & Doğan, 2010).

As a result of the research, it was discovered that there were differences with regard to the number of DIF items identified by three methods and the level of DIF that the items contained; however, methods were consistent in detecting uniform DIF. Some research also showed that MH and SIBTEST results comply with each other (Akin Arıkan, Uğurlu, & Atar, 2016; Narayanan & Swaminathan, 1994; Roussos & Stout, 1996).

It should be noted that there may be some advantages and disadvantages when DIF methods used in the research are examined in respect to the length of subtests. As subtests consist of 33, 15 and 12 items, they can be regarded as short tests. In their simulation study, Atalay Kabasakal, Arsan, Gök and Kelecioğlu (2014) reported that MH method had lower type I error in short tests compared with long tests and type I error with SIBTEST method increased when the length of tests decreased. From this point of view, in this research test length might have a positive impact on analyses with MH method and negative impact on analyses with SIBTEST method. Finch (2005) also reported that 20 items had an inflated type I error with MIMIC method with three parameter logistic data. However, it was discovered that 50 items had a lower type I error with three parameter logistic data. Therefore, DIF analyses with MIMIC method might be influenced negatively due to test length. Besides, Finch (2005) stated that as the size of focal group increased, power of SIBTEST and MH methods also increased. In this respect, focal group sizes might have positive impact on DIF analyses. Atalay Kabasakal, Arsan, Gök and Kelecioğlu (2014) found out that when the sizes of focal and reference groups were not equal, type I error was lower with MH method. Further, between the groups with different standard deviations, SIBTEST method had a lower type I error. In this research, the size of focal and reference groups was not equal, and there were differences between standard deviations, which may have contributed to DIF analyses.

DIF results showed that one item in translation subtest contained DIF in favor of male students. There were nine DIF items in vocabulary and grammar knowledge subtest, six DIF items in reading comprehension subtest and four DIF items in translation subtest in terms of school type. The reason why the number of DIF items by school type was higher than the number of DIF items by gender might be the serious gap between schools. Berberoğlu and Kalender (2005) investigated student achievement in Student Selection Examination (SSE) and The Programme of International Student Assessment (PISA) across years, school types and regions. It was found that student achievement changed dramatically according to school types because there is a notable difference in learning between school types. It is also supported by studies that there is a big gap between school types in Turkey (Arga, 2017; Yalçın, 2011; Yiğit, 2010). Research findings reveal the necessity to investigate the factors that cause differences between the school types and to take measures to reduce this difference.

Another finding of the study is that SIBTEST and MIMIC methods were more consistent in DBF analyses by gender compared with DBF analyses by school type. Finch (2012) noted that if group means are different on latent trait MIMIC method worked better than SIBTEST method. Therefore, the differences between methods in DBF analyses according to school type may have been caused by mean differences. Moreover, another reason for the inconsistency in DBF analyses in some item bundles might be the testing only uniform DBF with MIMIC method.

In addition, experts stated that one item showing DIF in terms of gender in translation subtest was not biased and evidence of bias was found in thirteen of nineteen items that contained DIF in terms of school type. Expert opinions also revealed that four of the seven items favoring AHS were grammar items which require knowledge. According to experts, being familiar with that



type of questions may become an advantage for AHS students. Socioeconomic status, scientific terms and education based on practice and speaking were suggested as the sources of bias for items favoring PHS. Bakan Kalaycıoğlu (2008) also reported that grammar items based on knowledge were in favor of AHS and items based on reading which do not require knowledge were in favor of PHS. Evidence of bias was found in items favoring VHS or RHVS due to expressions which students from these schools might be more familiar.

In this research, DIF and DBF analyses of English test in 2016 UPE were carried out with respect to gender and school type. Expert opinions were consulted to identify possible source of bias in items showing DIF. Student Selection and Placement Center carries out different language test every year, DIF analyses for these tests can be performed and a pattern for language tests may be formed in terms of bias sources.

Testlets which are frequently used as reading comprehension items in language tests might be examined in terms of DIF. Influence of different booklets on DIF can be studied as well.

## ORCID

Rabia Akcan  <https://orcid.org/0000-0003-3025-774X>

Kübra Atalay Kabasakal  <https://orcid.org/0000-0002-3580-5568>

## 5. REFERENCES

- Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, 24(1), 7-36. DOI: 10.1177/0265532207071510
- Akın Arıkan, Ç., Uğurlu, S., & Atar, B. (2016). A DIF and bias study by using MIMIC, SIBTEST, Logistic Regression and Mantel-Haenszel methods. *Hacettepe University Journal of Education*, 31(1), 34-52. DOI:10.16986/HUJE.2015014226
- Arga, B. (2017). Gender and student achievement in Turkey: School types and regional differences based on PISA 2012 data (Master's Thesis). İhsan Doğramacı Bilkent University, Ankara.
- Atalay Kabasakal, K., Arsan, N., Gök, B., & Kelecioğlu, H. (2014). Comparing performances (Type I error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel methods in the determination of differential item functioning. *Educational Sciences: Theory & Practice*, 14(6), 2186-2193. DOI: 10.12738/estp.2014.6.2165
- Bakan Kalaycıoğlu, D. (2008). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi [Item bias analysis of the University Entrance Examination]. (Doctoral Dissertation). Hacettepe University, Ankara.
- Berberoğlu, G., & Kalender, İ. (2005). Öğrenci başarısının yıllara, okul türlerine, bölgelere göre incelenmesi: ÖSS ve PISA analizi [Investigation of student achievement across years, school types and regions: The SSE and PISA analyses]. *Eğitim Bilimleri ve Uygulama*, 4(7), 21-35.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. London Sage.
- Chalmers, R. P. (2017). Improving the crossing-SIBTEST statistic for detecting non-uniform DIF. *Psychometrika*. DOI: 10.1007/s11336-017-9583-8
- Chalmers, R. P. (2018). *mirt, version 1.27.1: Multidimensional item response theory*. Retrieved from <https://cran.r-project.org/web/packages/mirt/index.html>
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement Issues and Practice*, 17(1), 31-44.
- Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, 33(4), 465-484.
- Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-



- Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, 29(4), 278-295. DOI: 10.1177/0146621605275728
- Finch, H. (2012). The MIMIC model as a tool for differential bundle functioning detection. *Applied Psychological Measurement*, 36(1), 40-59. DOI: 10.1177/0146621611432863
- Finch, H. W., & French, B. F. (2007). Detection of crossing differential item functioning a comparison of four methods. *Educational and Psychological Measurement*, 67(4), 565-582. DOI: 10.1177/0013164406296975
- Fox, J. (2016). *polycor, version 0.7-9: Polychoric and polyserial correlations*. Retrieved from <https://cran.r-project.org/web/packages/polycor/index.html>
- Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement*, 20(2), 26-36.
- Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A Confirmatory analysis. *Journal of Educational Measurement*, 38(2), 164-187.
- Gök, B., Kelecioğlu, H., & Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel-Haenszel ve Lojistik Regresyon tekniklerinin karşılaştırılması [The comparison of Mantel Haenszel and Logistic Regression techniques in determining the differential item functioning]. *Eğitim ve Bilim*, 35(156).
- Hallquist, M., & Wiley, J. (2018). *MplusAutomation, version 0.7-2: An R package for facilitating large-scale latent variable analyses in Mplus*. Retrieved from <https://cran.r-project.org/web/packages/MplusAutomation/index.html>
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun, *Test Validity* (pp. 129-145). Hillsdale NJ: Erlbaum.
- Kan, A. (2007). Test yansızlığı: H.Ü. Yabancı dil muafiyet sınavının cinsiyete ve bölümlere göre DMF analizi [Test fairness: DIF analysis across gender and department of H.U foreign language proficiency examination]. *Eurasian Journal of Educational Research*(29), 45-58.
- Karakaya, İ. & Kutlu, Ö. (2012). Seviye belirleme sınavındaki Türkçe alt testlerinin madde yanlılığının incelenmesi [An investigation of item bias in Turkish sub tests in Level Determination Exam]. *Eğitim ve Bilim* 37(165).
- Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, 61(4), 647-677.
- Lin, J., & Wu, F. (2003). Differential performance by gender in foreign language testing. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*.
- Magis, D., Beland, S., & Raiche, G. (2016). *difR, version 4.7: Collection of methods to detect dichotomous differential item functioning (DIF)*. Retrieved from <https://cran.r-project.org/web/packages/difR/index.html>
- Mcnamara, T., & Roever, C. (2006). Psychometric approaches to fairness: Bias and DIF. *Language Learning*, 56(S2), 81-128.
- Muthen, L. K., & Muthen, B. O. (1998). *Mplus user's guide*. Los Angeles.
- Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-haenszel and Simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, 18(4), 315-328.
- Osterlind, S. J. (1983). *Test item bias*. Sage Publications, Inc.
- Raiche, G., & Magis, D. (2015). *nFactors, version 2.3.3: Parallel analysis and non graphical solutions to the Cattell*. Retrieved from <https://cran.r-project.org/web/packages/nFactors/index.html>
- Rosseel, Y. (2017). *lavaan, version 0.5-23.1097: Latent variable analysis*. Retrieved from

- <https://cran.r-project.org/web/packages/lavaan/index.html>
- Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, 33(2), 215-230.
- Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/ DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58(2), 159-194.
- Willse, J. T. (2018). *CTT,version 2.3.2: Classical test theory functions*. Retrieved from <https://cran.r-project.org/web/packages/CTT/index.html>
- Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with Multiple Indicator Multiple Cause models. *Applied Psychological Measurement*, 35(5), 339-361. DOI: 10.1177/0146621611405984
- Yalçın, S. (2011). Türk öğrencilerin PISA başarı düzeylerinin veri zarflama analizi ile yıllara göre karşılaştırılması [The comparison of Turkish students' PISA achievement levels in relation to years via data envelopment analysis]. (Master's Thesis). Ankara University, Ankara.
- Yiğit, S. (2010). PISA matematik alt test sorularına verilen cevapların bazı faktörlere göre incelenmesi (Kocaeli-Kartepe örneği) [The analysis of the answers to PISA maths subtest questions according to certain factors (Kocaeli-Kartepe case)]. (Master's Thesis). Sakarya University, Sakarya.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. P. W. Holland, & H. Wainer içinde, *Differential Item Functioning* (s. 337-347). Hillsdale NJ:Erlbaum.
- Zumbo, B. D. (1999). *A Handbook on the theory and methods of differential item functioning (DIF): Logistic Regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottawa on Directorate of Human Resources Research and Evaluation, Department of National Defense.

## The Post-Graduate Academic English Language Skills and the Language Skills Measured by the Iranian PhD Entrance Exam: A Test Reform and Curriculum Change

Shiela Kheirzadeh <sup>1</sup>, S. Susan Marandi <sup>2,\*</sup>, Mansoor Tavakoli <sup>3</sup>

<sup>1</sup> Assistant Professor of TEFL, Alzahra University, Tehran, Iran

<sup>2</sup> Associate Professor of TEFL, Alzahra University, Tehran, Iran

<sup>3</sup> Professor of TEFL, University of Isfahan, Isfahan, Iran

### ARTICLE HISTORY

Received: 15 October 2018

Revised: 11 December 2018

Accepted: 02 January 2019

### KEYWORDS

Congruence,  
Field-specialist informants,  
Language-specialist informants,  
Academic skills,  
Post-graduate students,  
PhD entrance exam

**Abstract:** To investigate the congruence between the requisite post-graduate academic language skills and the language skills measured by the General English section of the Iranian National PhD Entrance exam, field-specialist informants, language-specialist informants and post-graduate students were questioned. The informants' data were collected through interviews and the students' data were obtained through a language skills' questionnaire. The informants and students' data were analyzed through content analysis and frequency analysis, respectively. The informants acknowledged that all four language skills were crucial for academic success. Considering congruity, both groups of informants asserted that there was little congruity between the language skills measured by the exam and those of the academic context. Post-graduate students believed that the reading section of the exam did not match their academic needs; they also believed that a writing section should be added and that a listening section need not be included in the exam. The findings have some implications for a change in the curriculum preceding the exam.

## 1. INTRODUCTION

Tests are commonly used for the purpose of making decisions such as selecting the right person for a job, awarding a certificate and entering a higher level of education (Zahedi & Shamsaee, 2012). If the consequences and the decisions made by the testing program are serious and affect a large group of people, the test is called high-stakes. The exact identification of test purpose and use are principal prerequisites for test development, especially in the case of high-stakes tests. Thus test developers need to consider, “for what purpose the test will serve, what underlying construct will be applied, who will take the test, and how it will be used and by whom” (Spaan, 2006, p. 71). Therefore, test designers need to consult the stakeholders to decide on test content and specifications. In other words, they should decide what language skills should be measured by the test and whether they should be measured integratively or discretely

---

CONTACT: S. Susan Marandi ✉ [susanmarandi@alzahra.ac.ir](mailto:susanmarandi@alzahra.ac.ir) 📠 Associate Professor of TEFL, Alzahra University, Tehran, Iran

ISSN-e: 2148-7456 / © IJATE 2019

(Spaan, 2006). According to Ryan (2002), the criteria to decide about the test content and purpose comes from a variety of resources, including the construct itself, various stakeholders, and existing research. In order to produce a language test that appropriately measures the academic language needs of the students, a detailed and precise analysis of the construct needs to be carried out (Butler et al., 2004). As stated by Berendes et al. (2018), academic language is the language used to transfer and acquire knowledge whether in spoken academic settings or in school textbooks. Ryan (2002) also suggests consulting stakeholders to decide about the assessment interpretation or test use to assure that the test is measuring what the students know and need. These stakeholders are teachers, students, parents or anyone involved in the educational system.

What can be inferred from the above mentioned points is that there must be congruence between what a test is measuring and the needs of the test takers, especially in the case of high-stakes tests. If the test results are used to decide on whom to accept for further studies at a higher educational level, ensuring the congruence between the academic language needs of the students and the needs of the test is imperative.

### **1.1. English for Academic Purposes (EAP)**

English for Academic Purposes (EAP) was a terminology coined by Johns in 1974 (Hyland, 2006). EAP, according to Hyland, “attempts to offer systematic, locally managed, solution-oriented approaches that address the pervasive and endemic challenges posed by academic study to a diverse student body by focusing on student needs and discipline-specific communication skills” (p. 4). Scarcella (2003) believes that learners must develop an advanced level of proficiency in all language skills to have successful “discipline-specific communication.” According to him, academic English is not limited to the reading skill, and learners should be able to use words in written and spoken communications, as well.

According to Gottlieb and Ernst-Slavit (2013), academic language or academic English is a register, a variety of language for a special group of audience in a specific context. In other words, academic language is characterized by specific linguistic features, discourse features, grammatical structures and vocabulary for a specific discipline. Chamot and O’Malley (1994) define academic language as “the language that is used by teachers and students for the purpose of acquiring new knowledge and skills” (p. 40). This is a rather general definition of academic language; however, attempts have been made to define academic language in a narrower sense, focusing on language functions, register, use and cognitive difficulty (e.g., Scarcella, 2003; Solomon & Rhodes, 1995). Moreover, the recent focus has been on language specific to special academic contexts such as English for engineering, social sciences, etc. (Gottlieb, 2004).

According to Saville-Troike (1984) and Swales (1990), target academic English language tasks mainly consist of research projects, summaries, writing critiques, note-taking in lectures, and reading abstracts and reports, all of which have their own genres. Scarcella (2003), citing Swales (1990), points out that target academic English tasks include “reading abstracts, getting down the key ideas from lectures, and writing critiques, summaries, annotated bibliographies, reports, case studies, research projects, expository essays” (p. 9). Dudley-Evans and St John (1998, p. 41) list the following tasks as the core academic tasks:

- Listening to lectures
- Participating in supervisions, seminars and tutorials
- Reading textbooks, articles and other material
- Writing essays, examination answers, dissertations and reports

A huge portion of the university work, especially in an EFL (English as a Foreign Language) context, is devoted to reading. Therefore, many EFL learners gradually acquire advanced

reading skills despite the fact that they do not demonstrate the same progress in oral communication skills (Huang, 2006; Shafie & Nayan, 2011). Reading, as stated by Flowerdew and Peacock (2001) and Grabe and Stoller (2011), is composed of macro- and micro-skills. Using the existing knowledge or background schemata to decipher the new information and knowledge is an example of a macro-reading skill. As stated by Grabe (2009), readers form their interpretation of the text by focusing on their feeling about it and whether it relates or contradicts background knowledge. In other words, they integrate text information with other ideas developed from their background knowledge and interpret it based on such background knowledge. Generalization, recognition, and finding logical relations are examples of micro-reading skills. Word recognition is considered the most important factor in successful reading since it is not possible for readers to comprehend without being able to recognize words quickly and accurately and being sensitive to orthographic, phonological, and semantic usages (Grabe, 2009; Kuzborska, 2010)

Studies by Durkin (2004) and Reid, Kirkpatrick and Mulligan (1998) indicated that non-native English learners need to spend twice or three times longer than native speakers to finish reading a passage; therefore, forming the habits of critical reading for academic purposes is much harder.

Though Shelyakina (2010) and Grabe and Zhang (2013) highlight the strong connection between reading and writing tasks, writing is considered as a major problem for students (Leki & Carson, 1994; Rosenfeld et al., 2001; Zhu & Flaitz, 2005). Bridgeman and Carlson (1983) investigated the beliefs and practices of academic writing. Their main findings were (a) at the graduate level, writing is the prime skill; (b) even for the disciplines in which writing was not of great emphasis, students were required to have writing assignments from the first year of college; (c) the type of required writing skills varied from discipline to discipline; (d) the assessment criterion was discourse, rather than word or sentence, and finally, (e) non-native speakers had more problems at the word or sentence level compared with the natives. Canseco and Byrd (1989, p. 308) list “examinations, problems and assignments, projects, papers, case studies, reports and miscellaneous writing assignments” as the seven categories of writing assignments of college students. Finally, Saville-Troike (1984) concluded that “the language skill which is most likely to develop . . . [academic] competence is writing” (p. 217). To achieve academic success and to present the academic skills, quality writing is imperative. Besides writing, oral communication skills are of utmost importance.

Berman and Cheng’s (2001) needs analysis of graduate and undergraduate students indicated that academic oral communication skills such as presentation or class discussions are the most difficult for L2 students. Forey and Feng (2016) stressed the importance of “teaching interactive, interpersonal features that help speakers engage with their audiences” (p. 428). Cheng, Myles and Curtis (2004) found that L2 graduate students had the most problems in speaking and writing skills. Ostler (1980) and Morell (2007) state that ESL (English as a Second Language) students are in need of help in developing academic speaking abilities such as talking with their instructors; she proposes that, “graduate [ESL-EAP] classes might need to include one unit on preparing and giving talks and another on preparing for and participating in panel discussions” (p. 501).

Listening is the most challenging skill for L2 (Second Language) learners, even the highly proficient ones (Mason, 1995, as cited in Ferris & Tagg, 1996; Field, 2011). Olster’s (1980) participants stated that every day listening and conversations were much easier for them than the classroom context conversations. Furthermore, Kim (2006), asserts that students with high scores in TOEFL test are not necessarily skillful enough in academic listening skill; she claims that academic listening has its own specialties, totally different from those of everyday listening. And finally, according to Benson (1989, p. 441), what is necessary is “listening to



learn” as opposed to “listening to comprehend.” In other words, academic listening is listening for meaning and understanding, and not the mere decoding of the language (Rost, 2011).

## **1.2. English for Academic Purposes (EAP), assessment and curriculum**

In an applied field of study such as language testing, the outcome of research into actual assessment designs and uses should model development for theories of measurement of language skills and building theories and models of language learning and use in target situations such as academic contexts (Schmitt & Hamp-Lyons, 2015). Though local EAP assessment, compared with standardized exams such as TOEFL, is an attempt to create a balance between assessment and the specific context, it can never fully represent what Bachman and Palmer (1996) call the target language use (TLU) domain. But since it has a closer connection with EAP teaching in that specific TLU context, it has the potential to better represent the construct of academic language skills. Therefore, the EAP and academic language testing communities can improve their understanding of students’ learning needs and the construct of EAP through cooperation in designing and developing EAP tests.

As stated by Schmitt and Hamp-Lyons (2015), if EAP assessments are to be used to make high-stakes decisions about candidates' readiness to progress to university or to graduate, it is important for those designing and developing such tests to arrive at the right level of authenticity; that is, “the degree of correspondence of the characteristics of a given language test task to the features of the TLU task” (Bachman & Palmer, 1996, p. 23). What can be implied from the definition of authenticity is that what students need at TLU domain, including language skills and abilities, should be reflected in the test. However, a point worth mentioning here is that the practical considerations of developing and administering large scale standardized proficiency exams or local ones mean that these are by their very nature reductionist; they narrow the educational curriculum. Curriculum narrowing (Nichols & Berliner, 2005, 2007; Watanabe, 2007) means limiting the educational curriculum through prioritizing some skills or abilities at the expense of others; for example, in the case of the General English section of the PhD Entrance exam in Iran, the academic reading skill is prioritized, which may underrepresent the construct of academic skills.

Therefore, the level, importance and difficulty of different target language academic skills crucial for academic success may vary, and carrying out a situation-specific study may help ensure the congruence between academic task and test task (Educational Testing service, 1990). The present paper reports the findings of a study which focuses on the following research questions:

1. Which academic English language skills are considered by field-specialist and language-specialist informants to be important for post-graduate students? And is there any congruence between these viewpoints and the language skills actually measured by the General English section of the Iranian National PhD Entrance exam?
2. Which academic English language skills do post-graduate students consider to be important? And is there any congruence between these viewpoints and the language skills actually measured by the General English section of the Iranian National PhD Entrance exam?

Before proceeding to the method section, a brief description of the PhD Entrance exam, the General English section, as the main concern of the present study, seems crucial. The General English section includes Grammar, Vocabulary and Reading sub-sections respectively. In the reading part, different reading skills such as guessing meaning from context, inferencing, finding main ideas, finding supporting details and making generalizations are measured. It should be mentioned that this exam is annually held and all the PhD candidates in all disciplines sit for it, though the post-graduate courses in Iran are commonly held in Persian and English is mainly used for supplementing the readings.



## **2. METHOD**

A total number of 183 participants took part in the data collection phase of the study. These participants were divided into three groups: The first group consisted of twelve field-specialist informants (faculty members of Humanities, Basic Sciences and Engineering departments) in different academic disciplines of Humanities (namely, History, Persian Literature, Geography and Psychology) (n=4), Basic Sciences (namely, Mathematics and Statistics, Chemistry, Biology and Physics) (n=4) and Engineering (namely, Computer Software, Electronic, Information Technology and Mechanics) (n=4) who were selected through purposive sampling from four public universities in Iran. The field specialist informants (as well as the language specialist informants) were selected from among those who were familiar with the Iranian National PhD Entrance Exam and had the experience of working with post-graduate students.

The second group consisted of five language specialist informants, meaning faculty members of English language departments. They were selected through purposive sampling from four public universities in Iran. The criteria for their selection were familiarity with the General English section of the Iranian National PhD Entrance Exam and experience in teaching at the post-graduate level.

The third and last group consisted of 166 post-graduate (PhD level) students (male and female) of seven major public universities in Iran, with an age range of 25 to 50, and majoring in Humanities (n = 86), Basic Sciences (n = 44) and Engineering (n = 36). The post-graduate participants were those who had passed the PhD Entrance Exam and had been accepted by one of the public universities in Iran to further their studies at the PhD level. To ensure the generalizability of the findings, the participants were randomly selected from public universities in Iran in the three major academic disciplines including Humanities, Basic Sciences and Engineering.

### **2.1. Instruments**

#### *Interview questions*

Five researcher-made interview questions were posed to check the congruence between the academic language skills needed for post-graduate students and those required for the General English section of the Iranian National PhD Entrance Exam (See Appendix A). These questions sought to uncover which language skills that are more important in different disciplines; whether these skills match the skills measured by the PhD Entrance exam; the reason of not including some of the skills, if any, in the exam, from their viewpoint, and finally whether the degree of the importance of skills differs across the three major academic disciplines, namely, Humanities, Basic Sciences and Engineering. The interview questions revolved around the four main language skills were developed referring to the literature, namely Saville-Troike (1984), Swales (1990), Dudley-Evans and St. John (1998) and Scarcella (2003) and were piloted on a similar group and revisions were made in the wording and order of presentation. The interviews were administered in Persian and were recorded for further analysis.

#### *Language skills questionnaire*

A questionnaire with 3 multiple-choice items and 8 open-ended items was designed to discover the post-graduate students' opinions about the academic language skills they need (six items) and the ones measured by the General English section of the Iranian National PhD Entrance Exam (five items) (See Appendix B). It should be mentioned that the questionnaire was constructed based on the literature on different definitions of academic English needs and skills proposed by scholars such as Saville-Troike (1984), Swales (1990), Dudley-Evans and St. John (1998) and Scarcella (2003). The questionnaire was developed in the post-graduate students' first language, Persian and was piloted on a similar group of participants.

## 2.2. Procedure

For the interview section of the research, both language specialist and field specialist informants were informed about the purpose of the research and their consent for participation was obtained. As stated above, the criterion for selecting these informants was their familiarity with the PhD Entrance Exam and the experience of teaching and working with post-graduate students in different disciplines. The face-to-face interviews were carried out in the language specialist and field specialist informants' offices in university and were recorded for further analysis. After the data collection, all the interviews were transcribed and translated into English to find answers to the questions. The questionnaire was distributed manually or through e-mail, and the participants were promised the confidentiality of the responses.

## 2.3. Data analysis

The questionnaire data were analyzed through frequency analysis conducted by SPSS, version 23 and the interview data were analyzed by content analysis in terms of codes and emerging themes.

## 3. RESULTS

The first research question of the study sought the important academic language skills at post-graduate level, and whether those skills have been appropriately measured by the Iranian National PhD Entrance Exam, the General English section.

The following are the ideas of the field-specialist and language-specialist informants about language skill(s) that are crucially important for a post-graduate student. The main points stated by these specialists are as follows.

- *"The four language skills are of equal importance and none can be considered as the most significant. They can be compared to the four pillars of a firmly-built house: the removal of any one would undoubtedly lead to the collapse of the house."* (Humanities specialist)
- *"Post-graduate students need to be able to read field-specific books in English and write their papers in English to share their findings with their academic community."* (Humanities specialist)

*"The most important skill at the post-graduate level is reading; however, there is no need for listening in the Iranian context."* (Humanities specialist)

- *"All skills are important; however, speaking, listening and writing are more significant, respectively. Nearly all PhD students are conversant with reading; therefore, there is no need to measure reading in the PhD Entrance Exam."* (Basic Sciences specialist)
- *"Reading and writing are the most notable ones; however, students need to be able to present in classes or conferences, which requires a good command of oral communication skills."* (Basic Sciences specialist)
- *"Reading and writing along with a good command of grammar and field-specific terminologies are imperative."* (Engineering specialist)
- *"Reading and writing have the highest priority, respectively. Listening and speaking are necessary only if PhD students want to take sabbatical leave abroad."* (Engineering specialist)
- *"Reading and then writing are more important. A PhD student is expected to read and write professionally in his own field."* (Language specialist)
- *"Reading, writing and the proper mastery of vocabulary, especially lexical bundles, is imperative."* (Language specialist)

The other issue of interest was whether the academic language needs of PhD students in the three abovementioned disciplines differ. The field-specialist informants, except for two, mentioned that there was no difference among the disciplines considering their academic needs. In other words, all four skills are important to all academic disciplines. Of those two informants who disagreed, one of them (from an Engineering background) stated that Engineering students feel a stronger need for English as they are dealing with technology and empirical sciences. The other informant (with a Humanities background) referred to the effect of culture on the humanities, which makes writing in a foreign language harder. In other words, transferring culture-bound thoughts and beliefs truly and exactly to another language with no common cultural background was felt to be more complicated than transferring information in a culture-free field of study such as Engineering.

The language-specialist informants, in general, believed that there was no substantial difference in the English language needs of the various academic disciplines and how they should prioritize the language skills across the three academic disciplines; at the same time, however, they claimed that Humanities students need to have a wider scope of discourse knowledge and that the academic language of the Engineering group is more symbolic and concrete while the language of the Humanities is more abstract and conceptual.

The last question was to find the field-specialist and language-specialist informants' opinions concerning the congruence between the skills measured by the General English section of the PhD Entrance Exam, mentioned above, and the skills that are essential for success at the PhD level. The opinions of the three groups of the informants are as follows.

- *"There is no congruence between the skills measured by the exam and the skills that are crucially important for PhD candidates. The reading passages and the vocabulary section have no bearings on their academic needs. Writing is of vital importance to a PhD candidate; however, it is not included in the exam. The reason might be that we should not test something in which we haven't invested time and effort teaching it to graduate or under-graduate students."* (Humanities specialist)
- *"There is no congruence between the skills measured by the exam and the skills that are crucially important for PhD candidates. There is no academic justification or theory to support the exam design and specification."* (Humanities specialist)
- *"PhD candidates really need writing skills; however, this skill is not included in the exam because of the scoring complexity. In general, the exam does not match the academic needs of PhD candidates."* (Basic Sciences specialist)
- *"The exam is not standard. It can be replaced by standard exams such as academic IELTS. The PhD students who are accepted in universities based on this exam are not competent in English."* (Basic Sciences specialist)
- *"There is no congruence between the exam and learners' academic needs. It is just an imitation of world-famous proficiency tests. The reading section is acceptable to some extent, but not that relevant. It is highly recommended to add writing and listening skills to the exam."* (Engineering specialist)
- *"The exam needs to be improved. A writing section must definitely be included in the exam, even if it is costly and time consuming to score and administer."* (Engineering specialist)
- *"The exam focuses only on reading and vocabulary, and this does not match the academic needs. The reason might be that measuring other language skills like writing and listening is difficult if not impossible."* (Language specialist)
- *"The focus is on language components (grammar and vocabulary). The subordinate role given to the essential language skills might be due to the design and evaluation complexities. However, excluding the language skills renders the exam invalid."* (Language specialist)

### **3.1. Post-graduate students**

The second research question of the present study explored the post-graduate students' own opinions about the academic language skills that are crucial to their academic success, and whether these skills were included in the General English section of the PhD Entrance Exam. Six out of the eleven items in the questionnaire dealt with the academic language skills, the results of which are presented below.

The first item of the questionnaire asked the post-graduate students to rank the four language skills based on their importance at the PhD level. Based on the result of the ranking, checked by the Friedman Test, the mean ranks indicated that reading (1.84), writing (2.31), listening (2.91) and speaking (2.94) were ranked from the most important to the least important, respectively.

The second item asked the students to indicate their most frequent use of the writing skill. The results indicated that writing articles (64.05%) was the most important writing need for Iranian post-graduate students. Writing emails and online correspondences with professors, researchers and post-graduate students abroad is the second most felt need (20.70%) followed by translating their papers and research findings from Persian into English (5.29%), writing theses and proposals in English (5.29%), report writing (2.88%) and preparing PowerPoint slides for classroom and conference presentations (2.40%). The third questionnaire item sought to uncover their most frequent use of the reading skill. As may be seen in [Table 1](#), the most important academic reading need was considered to be reading scientific articles (48.81%) to learn about the recent trends in research findings. Reading books (35.32%), different scientific texts (6.34%), websites (5.59%), theses and abstracts (2.37%) and finally reading newspapers (1.18%) were ranked next, respectively.

The main academic listening needs of the post-graduate students were as follows. 25.25% of the post-graduate participants of the present study mentioned that they required this skill to benefit from lectures and conference presentations. However, the same percent (25.25%) of the participants stated that they do not need English listening comprehension in their current academic contexts. Listening to and understanding talks and conversations while traveling abroad for the sabbatical leave, for instance, were ranked as the third most frequent (19.77%). Understanding films and documentaries related to their own disciplines (18.32%), news (7.92%) and online courses and classes (3.46%) were ranked next, respectively. Finally, the post-graduate students were asked about their need for English speaking skills in their current context. The two main academic speaking needs which were of nearly equal importance to the participants of the present study were presenting in international conferences and seminars (36.68%), and conversing and interacting with professors, researchers and post-graduate students abroad (36.48%); however, 30.68% of the participants stated that they do not need speaking in their current academic life ([Table 1](#)).

The last item concerning the academic needs of the post-graduate students required the participants to rank the eight listed academic skills (See Appendix B) based on their importance. The result of the Friedman Test indicated that the mostly required academic skill was writing abstract followed by writing articles, academic presentation, note-taking in lectures, writing critiques and summaries. Discussion and writing reports were ranked as the least felt needs in the academic context.

While the first six items of the questionnaire investigated the academic language needs of PhD students, the next five items examined whether the current entrance exam was indeed congruent with these academic needs or not:

**Table 1.** *Post-Graduate Students' Viewpoints regarding Academic Skills' Needs*

| skill     | Item                                  | Percent (%) |
|-----------|---------------------------------------|-------------|
| Writing   | Articles                              | 64.05       |
|           | Emails and letters                    | 20.70       |
|           | Translations                          | 5.29        |
|           | Proposals and theses                  | 5.29        |
|           | Reports                               | 2.88        |
|           | Power point slides                    | 2.40        |
| Reading   | Articles                              | 48.81       |
|           | Books                                 | 35.32       |
|           | Texts                                 | 6.34        |
|           | Websites                              | 5.59        |
|           | Theses and abstracts                  | 2.37        |
|           | Newspapers                            | 1.18        |
| Listening | Lectures and conference presentations | 25.25       |
|           | No need                               | 25.25       |
|           | Talks and conversations               | 19.77       |
|           | Films and documentaries               | 18.32       |
|           | News                                  | 7.92        |
|           | Online courses and classes            | 3.46        |
| Speaking  | Conferences and seminar presentations | 34.68       |
|           | Conversation                          | 34.48       |
|           | No need                               | 30.68       |

The first of the five items inquired about the necessity of a writing skill component (which is non-existent in the current exam). The frequency analysis of the responses indicates that 65.71% of the participants preferred a writing section to be included in the exam while 34.29% did not. Those participants who favored the inclusion of this section to the exam asserted that,

- Writing articles is the most important need at the post-graduate level (43.47%)
- Writing is of vital importance in their academic life (11.38%)
- Writing is an instrument for staying in touch with the academic community (10.86%)

The participants who disapproved (34.29%) the inclusion of a writing section to the General English Section of the Iranian National PhD Entrance Exam reasoned that:

- The educational system neither considers this skill as a priority in teaching nor is it included in the higher education syllabuses; therefore, it should not be tested (12.15%)
- Writing is not an academic need, since they can get help from others to write (12.06%)
- The PhD Entrance exam is not the proper place to measure this skill (6.05%)
- Grammar and vocabulary can account for the writing skill (4.03%)

Nearly fifty percent (49.70%) of the participants stated that the reading comprehension section of the exam matches their academic needs, while 50.30% were not satisfied with this section of the exam due to the following reasons:

- The reading passages are not field-specific so they do not match academic needs (25.92%)
- A multiple-choice test is not a valid measure of the academic reading skill (15.70%)
- The reading passages are difficult (8.68%)

Similar to writing, listening comprehension is also currently excluded from the General English section of the Iranian National PhD Entrance Exam. Only 23.50% of the participants



requested the addition of a listening component, while 76.50% did not deem this to be necessary. Those who favored the continued exclusion of listening from the exam commented that:

- There is no need for the English listening skill in the [Iranian] academic context (42.37%)
- Listening comprehension is not a priority compared with other skills (25.59%)
- There are not enough proper facilities to measure this skill during the exam session (4.44%)
- The listening comprehension section is a source of stress in the exam (4.10%)

Speaking is also currently excluded from the General English section of the Iranian National PhD Entrance Exam. On being asked for an opinion, 70.52% of the students claimed there is no need to include a speaking section to the PhD Entrance, whereas 28.95% considered the addition of a speaking section to the exam as compulsory. Those who were against the inclusion of this skill (70.52%) reasoned as follows:

- There is no need for the English speaking skill in the [Iranian] academic context (31.07%).
- Though it is an important skill, it is not a priority compared with other skills (27.14%)
- The subjective judgment of the raters may affect the scores (12.31%)

The most popular reasons proposed by those who were in favor of the addition of a speaking section to the exam (28.95%) were as follow:

- Speaking facilitates interaction with colleagues and other students worldwide (13.79%)
- Speaking is one of the main language skills (11.59%)
- Speaking helps strengthen our field-specific knowledge (3.57%)

The last item in the questionnaire inquired about the degree of satisfaction of the students with the total test from an academic needs standpoint. 13.9% post-graduate participants of the study were satisfied with the exam, 40.4% believed that the exam did not match their academic needs. 34.3% considered the exam acceptable to some extent and finally, 10.8% were undecided.

#### **4. DISCUSSION and CONCLUSION**

The purpose of the present study was to discover if the language skills measured by the General English section of the Iranian National PhD Entrance Exam corresponded to the target language academic skills required at the post-graduate level in Iran. For this purpose, a data triangulation method was selected in which the data were collected from field-specialist informants in the academic disciplines of Humanities, Basic Sciences and Engineering, language-specialist informants, and the post-graduate students in the three mentioned disciplines in Iranian public universities.

The first research question was to find the field-specialists' opinions about the language skills that are most important at the post-graduate level and whether there was congruence between the exam and the required academic skills. The general conclusion from the field-specialist informants' opinions was that all language skills (listening, speaking, reading and writing) were imperative for post-graduate students. This conclusion is in line with Scarcella (2003) who believes that learners must develop an advanced level of proficiency in all language skills to have successful discipline-specific communication. However, reading and writing skills were considered as the most important skills at the post-graduate level, which is in line with Ostler's (1980) study who found that reading was the first important consideration in the academic



context, and is also in sync with Saville-Troike (1984), who claimed that “the language skill which is most likely to develop. . . [academic] competence is writing” (p. 217). All in all, the field-specialist informants believed that the current exam does not meet the expectations of the post-graduate level. This is in line with what has been stated by Atai (2002a), who investigated the curriculum development of English for Specific Academic Purposes (ESAP) in Iran and concluded that, “ESAP curriculum development in Iran has not been conducted systematically and coherently.... the participants involved in the development and implementation of ESAP programs have typically done their tasks independently of each other” (p. 1). This has led to some participants suggesting the revision or the replacement of the current exam with international worldwide exams such as Academic IELTS.

The first research question also dealt with language-specialists. In general, the language specialists stated that all four language skills were imperative for a post-graduate student; however, in the academic context of Iran, reading and writing are more important. This incongruence might be justified by what has been proposed by Atai (2002b) who argues that the lack of rhetoric between the upper and lower layers of the English for academic purposes (EAP) curriculum results in confusion in EAP courses in Iran. Therefore, two language skills, as stated by language specialists are prioritized at the expense of others. Furthermore, language specialists believed that there was no difference among the Humanities, Basic Sciences and Engineering students in their academic language needs; however, in assigning the specifications of the skills, the genre and discourse specialties of different disciplines should be highlighted for students especially in reading their field-specific passages or writing abstracts or articles. This finding is in line with Calson (1983) who found that the type of required writing skills varies from discipline to discipline and the assessment criterion should be discourse rather than word or sentence. Moreover, Shih (1992) suggests that the academic writing skill involves not only formal schemata but also content schemata. With regard to congruity, the language-specialist informants considered the test as a mismatch with the academic needs which, in turn, has rendered the test invalid. Kiany Mirhosseini, and Navidinia (2011) examined Iranian national documents to see if the literature on foreign language education policy has been taken into consideration in developing these documents. They concluded that the documents did not appear to be articulating coherent policies, and that there are occasional “mismatches among these documents” (p. 63). It is natural to expect that such mismatches would lead to similar problems in other areas as well, such as local assessment.

The second research question inquired about the language skills and congruity from the perspective of the post-graduate students. The order of the importance of skills was reading, writing, listening and speaking. Writing was mostly required for writing articles and email correspondences with researchers, professors and post-graduate students abroad. Reading skill is important for reading articles and books. Considering the listening skill, a successful and comprehensible presentation in conferences and seminars was the speaking need of post-graduate students.

As mentioned above, the General English section of the PhD Entrance Exam does not currently include writing, listening or speaking skills. The only measured skill is the reading comprehension skill. Therefore, this study also aimed at investigating if post-graduate students felt the need for the inclusion of these excluded skills, and whether the existing reading test matched their target academic needs. A large number of the participants requested the inclusion of writing skill to the exam since they need to write articles and to be in touch with their own academic community. The main reason for those who disapproved the inclusion of writing skill was that it was unfair to include a skill in the exam which had not been taught or prioritized in the academic context; therefore, it should not be tested. Students’ claim might be justified through ‘personnel policy’ lens which examines whether there are language proficiency

standards for teachers in academic contexts (Baldauf et al., 2010), and since in Iran, there is no systematic definition in the form of a formal evaluation scheme dealing with teachers' language proficiency standards or their professional knowledge of TLU tasks (Atai & Mazlum, 2013), therefore writing skill is totally neglected. Furthermore, what has been stated by the post-graduate students regarding the exclusion of writing in the academic context of Iran has previously been mentioned by Eslami-Rasekh and Valizadeh (2004) and Farhady and Hadayati (2009) who state that though students show a great interest in communication skills, the grammar-translation, traditional and form-focused educational system with a great emphasis on grammar and translation has lessened their chance of using the language for communicative purposes. Furthermore, according to Clapham (2000), merely testing grammar is insufficient, and tests need to include writing tasks that are representative of the academic tasks. Considering the high percent of the students asking for the inclusion of a writing task, it is recommended to add a writing section to the future administrations of the exam. The type of writing tasks, as proposed by field and language specialist informants, can be similar to the IELTS academic exam.

The majority of the participants (76%) preferred the exam not to have a listening comprehension section since they neither needed listening in their academic context, nor was it a priority comparing with other skills; also it was considered to be stressful. Similar to the listening skill, 70.52% of the participants were not in favor of a speaking section in the PhD Entrance Exam, reasoning that they did not need it and that the subjective judgment of the raters might affect their scores.

As stated before, reading comprehension is the only skill included in the General English section of the Iranian National PhD Entrance Exam; however, approximately half the participants claimed that the existing reading test does not match their academic needs while the other half believed it did. Therefore, it is recommended that the test developers do a thorough revision of the reading section of the exam to match the expectations of post-graduate students. The test designers are recommended to consider students' ideas, among others, in their test design, as suggested by Fox and Cheng (2007), who believe that underrepresentation of test-takers' perspectives in language assessment contexts is clearly problematic. These scholars emphasize including validation evidence from test-takers, such as an analysis of "how test takers interpret test constructs and the interaction between these interpretations, test design, and accounts of classroom practice" (Fox & Cheng, 2007, p. 9).

### *Implications of the study: Test reform*

What can be generally inferred from these triangulated data is that the current Iranian PhD Entrance Exam does not fully match the language needs of post-graduate students, nor does the curriculum that precedes it adequately prepare the students for target language academic English needs at the doctoral level. It would therefore seem that at least some faulty decisions lie with the curriculum behind the exam which might need deeper consideration or even change by the Ministry of Science, Research, and Technology, policy makers and the curriculum developers, as mentioned by PhD exam candidates regarding the writing skill, it seems totally unfair to include it in the exam though it is of utmost importance.

Tusi (1998) believes that one of the problems of mainstream ELT material developers in the ministries of education is that they simply do not identify learners' needs. In a similar vein in the Iranian context, Maftoon et al. (2010, p. 2) argue that, "curriculum developers. . . have almost certainly neglected to pay attention to students' needs and future demands." Atai and Mazlum (2013) believe that the gap between planning and practice levels is the result of a centralized policymaking approach of the Iranian officials in the Ministry of Research, Science and Technology.

Fortunately, positive developments have recently been taking place in educational system of Iran, considering teaching English, to meet the target language academic needs of the students, such as changing the high school English books (which might be the root of the problem) as well as going over the old English teaching curricula at graduate and post graduate levels at the Ministry of Education and Ministry of Research, Science and Technology.

## ORCID

Shiela Kheirzadeh  <https://orcid.org/0000-0003-4665-0554>

S. Susan Marandi  <https://orcid.org/0000-0001-9852-1880>

Mansoor Tavakoli  <https://orcid.org/0000-0002-4029-466X>

## 5. REFERENCES

- Atai, M.R. (2002a). ESAP curriculum development in Iran: An incoherent educational experience. *Journal of Persian Literature and Human Sciences of Tehran Teacher Training University*, 1, 17–34.
- Atai, M.R. (2002b). Iranian EAP programs in practice: A study of key methodological aspects. *Sheikhbahaee Research Bulletin* 1(2), 1–15.
- Atai, M. R., & Mazlum, F. English language teaching curriculum in Iran: Planning and practice. *The Curriculum Journal* 24(3), 389-411.
- Baldauf, R.B., Li, M., & Zhao, Sh. (2010). *Language acquisition management inside and outside the school*. In *Handbook of educational linguistics*. Oxford, UK: Blackwell.
- Berendes, K., Vajjala, S., Meurers, D., Bryant, D., Wagner, W., Chinkina, M. (2018). Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*, 110, 518–543.
- Benson, M. J. (1989). The academic listening task: A case study. *TESOL Quarterly*, 23(3), 421-445.
- Berman, R., & Cheng, L. (2001). English academic language skills: Perceived difficulties by undergraduate and graduate students, and their academic achievement. *Canadian Journal of Applied Linguistics*, 4, 25-40.
- Bridgeman, B., & Carlson, S. B. (1983). *Survey of academic writing tasks required of graduate and undergraduate foreign students*. (TOEFL Research Report No. 15). Princeton, NJ: Educational Testing Service.
- Butler, F. A., Lord, C., Stevens, R., Borrego, M., & Bailey, A. L. (2004). *An Approach to Operationalizing Academic Language for Language Test Development Purposes: Evidence from Fifth-Grade Science and Math*. CSE Report 626. US Department of Education.
- Canseco, G., & Byrd, P. (1989). Writing required in graduate courses in business administration. *TESOL Quarterly*, 23(2), 305-316.
- Chamot, A. U., & O'Malley, J. M. (1994). *The CALLA handbook: Implementing the cognitive academic language learning approach*. Reading, MA: Addison-Wesley.
- Cheng, L., Myles, J., & Curtis, A. (2004). Targeting language support for non-native English speaking graduate students at a Canadian university. *TESL Canada Journal*, 22, 50-71.
- Clapham, C. (2000). Assessment for academic purposes: where next? *System*, 28(4), 511-521.
- Dudley-Evans, T., & St John, M.J. (1998). *Developments in English for Specific Purposes*. Cambridge: Cambridge University Press.
- Durkin, K. (2004). *Challenges Chinese students face in adapting to academic expectations and teaching/learning styles of UK Masters courses, and how cross cultural understanding*

- and adequate support might aid them to adapt. Discussion Paper. London: British Council.
- Educational Testing Service (1990). *TOEFL test and score manual*. Princeton, NJ.
- Eslami-Rasekh, Z. & Valizadeh, K. (2004). Classroom activities viewed from different perspectives: Learners' voice vs. teachers' voice. *TESL EJ*, 8(3), 1-13.
- Farhady, H. & H. Hedayati. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics*, 29, 132-141.
- Ferris, D., & Tagg, T. (1996). Academic listening/speaking tasks for ESL students: Problems, suggestions, and implications. *TESOL Quarterly*, 30(2), 297-320.
- Field, J. (2011). Into the mind of the academic listener. *Journal of English for Academic Purposes*, 10, 102-112.
- Flowerdew, J., & Peacock, M., (2001). *Research perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press.
- Forey, G., & Feng, D. (2016). Interpersonal meaning and audience engagement in academic presentations: A multimodal discourse analysis perspective. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 416–30). New York, NY: Routledge.
- Fox, J., & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers. *Assessment in Education: Principles, Policy and Practice*, 14(1), 9–26.
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.
- Grabe, W., & Stoller, F. L. (2011). *Teaching and researching reading* (2nd ed.). London, UK: Pearson Education.
- Grabe, W., & Zhang, C. (2013). Reading and writing together: A critical component of English for academic purposes teaching and learning. *TESOL Journal*, 4(1), 9-24.
- Gottlieb, M. (2004). *Overview*. In WIDA consortium K -12 English language proficiency standards for English language learners: Frameworks for large-scale state and classroom assessment. Overview document. Madison: State of Wisconsin.
- Gottlieb, M. H., & Ernst-Slavit, G. (2013). Academic language a foundation for academic success in mathematics. In M. H. Gottlieb and G. Ernst-Slavit (Ed.), *Academic language in diverse classrooms: mathematics, grades K-2: Promoting content and language learning* (pp. 1-34). Corwin Press.
- Huang, S. C. (2006). Reading English for academic purposes—What situational factors may motivate learners to read? *System*, 34(3), 371-383.
- Hyland, K. (2006). *English for Academic Purposes: An Advanced Resource Book*. New York.
- Leki, I., & Carson, J. (1994). Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL Quarterly*, 28, 81–101.
- Kiany, Gh., Mirhosseini, S.A., & Navidinia, H. (2011). Foreign language education policies in Iran: Pivotal macro considerations. *Journal of English Language Teaching and Learning*, 53(222), 49–70.
- Kim, S. (2006). Academic oral communication needs of East Asian international graduate students in non-science and non-engineering fields. *English for Specific Purposes* 25, 479-489.
- Kuzborska, I. (2010). *The relationship between EFL teachers' beliefs and practices in reading instruction to advanced learners of English in a Lithuanian University context*. (Unpublished doctoral dissertation). University of Essex. Colchester, UK.

- Maftoon, P., M. Yazdani Moghaddam, H. Golebostan, & S.R. Beh-Afarin. (2010). Privatization of English education in Iran: A feasibility study. *The Electronic Journal for English as a Second Language*, 13(4), 1–12.
- Morell, T. (2007). What enhances EFL students' participation in lecture discourse? Student, lecturer and discourse perspectives. *Journal of English for Academic Purposes*, 6, 222–237.
- Ostler, S. E. (1980). A survey of academic needs for advanced ESL. *TESOL Quarterly*, 4(4), 489-502.
- Reid, I., Kirkpatrick, A., & Mulligan, D. (1998). *Framing reading*. Perth: National Center for English Language Teaching and Research with the Center for Literacy, Culture and Language Pedagogy at Curtin University of Technology.
- Rosenfeld, M., Leung, S., & Oltman, P. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels*. TOEFL monograph 21. Princeton, NJ: Educational Testing Service.
- Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Harlow, UK: Pearson
- Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice*, 21(1), 7-15.
- Saville-Troike, M. (1984). What really matters in second language learning for academic achievement? *TESOL Quarterly*, 18 (2), 199-219.
- Scarcella, R. (2003). *Academic English: A Conceptual Framework* (Technical report 2003-1). Santa Barbara, CA: Linguistic Minority Research Institute.
- Schmitt, D., & Hamp-Lyons, L. (2015). The need for EAP teacher knowledge in assessment. *Journal of English for Academic Purposes*, 18, 3-8.
- Shafie, L., & Nayan, S. (2011). The characteristics of struggling university readers and instructional approaches of academic reading in Malaysia. *International Journal of Human Sciences* [online]. 8, 1.
- Shelyakina, O. K. (2010). *Learner perceptions of their ESL training in preparation for university reading tasks* (Master's thesis). Brigham Young University – Provo.
- Shih, M. (1992). Beyond comprehension exercises in the ESL academic reading class. *TESOL Quarterly*, 26(2), 289-318.
- Solomon, J., & Rhodes, N. (1995). *Conceptualizing academic language* (Research Rep. No. 15). Santa Cruz: University of California, National Center for Research on Cultural Diversity and Second Language Learning.
- Spaan, M. (2006). Test and item specifications development. *Language Assessment Quarterly: An International Journal*, 3(1), 71-79.
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Zahedi, K., & Shamsaee, S. (2012). Viability of construct validity of the speaking modules of international language examinations (IELTS vs. TOEFL iBT): evidence from Iranian test-takers. *Educational Assessment, Evaluation and Accountability*, 24(3), 263-277.
- Zhu, W., & Flaitz, J. (2005). Using focus group methodology to understand international students' academic language needs: A comparison of perspectives. *TESL-EJ*, 8(4), 1-11.



**APPENDICES**

**Appendix A: Interview Questions**

1. Which language skills (listening, speaking, reading, writing and the two components of grammar and vocabulary) are required for PhD candidates? Why?
2. From among these skills, which is more important? Why?
3. Does the importance of the skills differ considering academic discipline (Humanities, Basic sciences, Engineering)? If yes, why?
4. On which skills the exam has focused? Why?

**Appendix B: Language Skills Questionnaire**

**The following questionnaire is designed to evaluate the English language needs of the non-English major PhD candidates. Please read the question and answer.**

- 1. Please rank the following language skills based on their order of importance (1 for the most important and 4 for the most important).**

Listening.....  
Speaking.....  
Reading.....  
Writing.....

- 2. For what do you use the writing skill? Please name.**

.....  
.....  
.....

- 3. Do you prefer the writing skill to be included in the PhD Entrance exam? Why?**

.....  
.....  
.....

- 4. For what do you use the reading skill? Please name.**

.....  
.....  
.....

- 5. Does the reading section of the PhD Entrance exam match your academic needs at PhD level? Why?**

.....  
.....  
.....

- 6. For what do you use the listening skill? Please name.**

.....  
.....  
.....



**7. Do you prefer the listening skill to be included in the PhD Entrance exam? Why?**

.....  
.....  
.....

**8. For what do you use the speaking skill? Please name.**

.....  
.....  
.....

**9. Do you prefer the speaking skill to be included in the PhD Entrance exam? Why?**

.....  
.....  
.....

**10. From among the English language skills required at the PhD level, the followings can be named. Please rank them based on their importance (from 1, as the most important to 8, as the least important) and add any other skill you may find necessary.**

- Reading abstracts and papers.....
- Taking notes in international conferences and seminars.....
- Presenting in international conferences.....
- Writing papers.....
- Writing a review on articles and books.....
- Summarizing.....
- Writing report.....
- Participating in scientific discussions.....
- Etc.....

**11. In general, does the English section of the PhD Entrance exam match your PhD level English needs?**

Yes..... No..... To some extent..... No idea.....

## Development of a Measurement Tool for Sustainable Development Awareness

Ayşe Ceren Atmaca<sup>1</sup>, Seyit Ahmet Kiray<sup>1,\*</sup>, Mustafa Pehlivan<sup>1</sup>

<sup>1</sup> Necmettin Erbakan University, Ahmet Kelesoglu Education Faculty, Department of Mathematics and Science Education, Division of Science Education, Konya/Turkey

### ARTICLE HISTORY

Received: 07 September 2018

Revised: 26 December 2018

Accepted: 18 January 2019

### KEYWORDS

Sustainability,  
Sustainable development,  
Sustainable development  
awareness

**Abstract:** Raising the sustainable development awareness is of great importance for the continuation of the world's livability. Teachers have a great responsibility in order for individuals to make sustainable development a part of their lives. For this reason, teachers need to be individuals with sustainable development awareness. In this study, it was aimed to develop a scale for determining the sustainable development awareness of teacher candidates. The developed scale consists of three sub-dimensions including economy, society, and environment and a total of 36 items. 425 science teacher candidates from seven state universities in Turkey participated in the research. The Cronbach's alpha reliability coefficient for the overall scale was calculated as 0.91. Confirmatory factor analysis was performed for the validity of the scale. In light of the analyses, the scale was found to possess the qualifications to determine the sustainable development awareness of science teacher candidates.

## 1. INTRODUCTION

Industrialization in the 20th century with the destruction of the environment caused by the unconscious steps taken in the name of development under the influence of rapid population growth together with urbanization brought concerns about human health and the future of the world (Altunbaş, 2003). In the aftermath of the destruction of natural life and the unconscious use of resources, many parts of the world have begun to suffer from food and water scarcity and, consequently, many deadly problems such as hunger, diseases, and poverty. In addition, climate change and global warming have become the most important issues affecting the future of our planet (Yerdelen, Cansiz, Cansiz, & Akcay, 2018). It has been noticed by all the communities that the Earth's self-renewal capacity is severely damaged. It has been understood that in the continuation of this course, the Earth will lose its ability to become a livable planet. Noticing that these problems will threaten not just a certain region but the whole world if not taken care of has led the search of solutions in the global scale (Baykal & Baykal, 2008). It has been decided that sustainable development should be included in education programs by

**CONTACT:** Seyit Ahmet Kiray ✉ [ahmetkiray@gmail.com](mailto:ahmetkiray@gmail.com) 📧 Necmettin Erbakan University, Ahmet Kelesoglu Education Faculty, Department of Mathematics and Science Education, Division of Science Education, Konya/Turkey

**ISSN-e: 2148-7456 /© IJATE 2019**

understanding that these problems can be solved only if every individual, every society, and every state living on the Earth are able to work together in collaboration and take certain responsibilities (Biasutti & Frate, 2017; Erten, 2015).

### **1.1. Sustainable Development**

Sustainable development was first officially discussed in the Brundtland Report published by the World Commission on Environment and Development in 1987, and the corresponding rapporteur defined it as “sustainable development that meets the needs of the present generation without compromising the ability of future generations to meet their own needs” (WCED, 1987). When looking at the definitions of sustainable development, international texts and sustainable development approaches, it is seen that sustainable development has three dimensions, namely economy, environment, and society (Borg, Gericke, Höglund, & Bergman, 2012; Olsson, Gericke, & Chang Rundgren, 2016). In order for sustainable development to take place, the sustainability of these three dimensions must be ensured simultaneously (Alkış, 2007; Sandel, Öhman, & Östman, 2006).

The society dimension of sustainable development includes the concepts of human rights, gender equity, peace and human security, cultural diversity and inter-cultural understanding (UNESCO, 2006), social services, health and education right, and social justice (Atmaca, Kiray, & Pehlivan, 2018; Özmete & Akgul-Gök, 2015).

Environmental sustainability includes issues such as the protection of natural resources (water, air, soil, energy, agriculture, and biodiversity), sustainable urbanization (UNESCO, 2006), reduction of environmental pollution (water, air, soil pollution), the use of renewable energy sources (geothermal, wind energy, etc.) instead of non-renewable energy sources (coal, petrol, etc.), protection of forests and increasing green areas, reduction of resource usage and environmental pollution by recycling of wastes, ecological footprint minimization, and stopping the global warming (Atmaca, et al., 2018; Koçak & Balcı, 2010).

Economic sustainability, on the other hand, includes issues such as conservative use of resources, income and expense balance, elimination of income distribution inequality, sustainable production and cost, reliable environments for investments, investments in high-income sectors, investments in vital sectors, and research and development (Atmaca, et al., 2018; Kuşat, 2013; Olsson, Gericke, & Chang Runghen, 2016; Şahin & Kutlu, 2014).

### **1.2. Scale Development Studies on Sustainable Development**

Türer (2010) developed the Sustainable Development Awareness Questionnaire as a measurement tool in his study that aimed to determine sustainable development awareness of social science and science teacher candidates. The questionnaire, developed by the researcher, consists of 3 sub-dimensions including social, economic and environmental in accordance with the theoretical framework of sustainable development in the literature and consists of 21 items. Chow and Chen (2012) developed a corporate sustainable development scale to determine how well the management strategies of companies overlap with the sustainable development contexts. Similar to Türer’s study, the developed scale consists of 22 items that are prepared for social, economic, and environmental sustainability contents frequently encountered as dimensions of sustainable development in the literature. The ‘Survey of Education for Sustainable Development Competencies (SSESDC)’ that was used by Biasutti and Surian (2012) in their study in order to investigate sustainable development competence areas of students from different universities was developed by a research team led by Professor Vassilios Makrakis in the RUCAS Tempus project. The survey includes an education dimension in addition to the environmental, economic and social dimensions that are frequently encountered as dimensions of sustainable development in the literature. The educational dimension includes

items in areas such as the attitudes towards sustainable development education, learning to exist, learning to live together in a sustainable way, learning to know, learning to do, learning to improve oneself and society. Biasutti and Frate (2017) developed a scale to be used to determine students' attitudes towards sustainable development in their study. The scale includes three dimensions (economy, society, environment) that are common in the literature, such as the measure developed by Biasutti and Surian (2012), with an addition of the educational dimension. In the dimension of education on the scale, areas such as student-centered teaching methods, future-oriented thinking, higher-order thinking skills, critical thinking, interdisciplinary, and related global and local subjects were emphasized. The sustainable development attitude scale developed after the validity and reliability analyzes was finalized with 4 sub-dimensions and 20 items. Kaya (2013) also developed a sustainable development attitude scale with the aim of determining the attitudes of secondary school students towards sustainable development. The developed scale consists of 3 sub-dimensions including social, environmental and economical and has 21 items in total. Similar to Biasutti and Frate's (2017) study, Manju (2015) developed a survey to investigate the views of teachers on the role of education in sustainable development. The developed questionnaire consists of 25 items focusing on sustainable development and its relationship to education. Unlike other studies, Doğan, Bulut ve Çımrın (2015) focused on sustainable consumption behaviors. Doğan, Bulut and Çımrın (2015) developed Sustainable Consumption Behavior Scale, which consists of 20 items, in their study that aimed to develop a scale for measuring sustainable consumption behaviors. The scale has four factors as environmental awareness, unnecessary purchase, saving, and reusability.

### **1.3. The Importance of Research**

The only way for sustainable development activities to reach its goal and become a way of life is to raise individuals who have sustainable development awareness and who shape their lives in the direction of sustainable development principles. The only way to raise individuals with sustainable development awareness is education (Aydoğan, 2010). Education is at the center of sustainable development. However, it is not meant to be an educational information accumulation. Education for sustainable development is aimed at educating individuals in accordance with sustainable development principles such as knowledge, attitudes, values, and behavior through a program that includes environmental, economic, and social issues (Summers, Kruger, Childs, & Mant, 2010).

If individuals are meant to be educated with sustainable development awareness, educators in all branches should be individuals with sustainable development awareness. In order for the education process to be carried out in an appropriate and productive way, it is necessary for teachers to have an awareness of their field and therefore to have sufficient knowledge, skills, values, and attitudes regarding that field. In this context, in order to educate individuals who have sustainable development awareness, teacher candidates who chose teaching as a profession and have been trained in education faculties should begin this profession as individuals with sustainable development awareness once they graduate (Demirbaş, 2015; Kahyaoğlu 2011).

Science teachers, one of the lessons expected to bring sustainable development skills to individuals, has a vital role. Science teachers need to be educators who have awareness of sustainable development in order to give students the objectives that will provide the opportunity to make sustainable development a way of life by teaching science lesson appropriately for its purpose. It is expected that today's science teacher candidates who will be future science teachers will have sustainable development awareness. This research is a scale development study designed to determine the sustainable development awareness levels of

science teacher candidates. When the scales containing the economic, social, and environmental dimensions in the literature are examined, it can be seen that the items of the economy and the society dimensions are different while the scales have almost similar properties in the items of the environmental dimension. This scale differs from other studies especially in terms of the approach to these two dimensions. Studies in the literature often have small number of items, which reduce the content validity of sustainable development sub-dimensions. In this study, a scale was developed which has almost twice the number of items in the other studies, in which the content determined with the help of the literature and experts' opinions was fully reflected to the scale items. That is why this scale also has a more inclusive feature than other scale development studies.

## **2. METHODS**

In this study, the validity and reliability analyses of the sustainable development awareness scale developed for science teacher candidates were conducted.

### **2.1. Study Group**

Sustainable Development Awareness Scale developed in the study was given to 425 science teacher candidates who were senior students at seven different universities in Turkey during the 2017-2018 academic year. The Departments of Science Education in Turkey are divided into three groups as lower, middle, and upper according to the university entrance scores. To apply the scale, the universities have been determined by choosing approximately the same number of pre-service teachers from these three groups such as 2 from the upper group, 2 from the middle group, and 3 from the lower group. The teacher candidates participating in the survey were given 20 minutes to answer the whole scale. Of the 425 science teachers who participated in the research, 340 (80%) were female and 85 (20%) were male.

### **2.2. Data Collection Tool Development Process**

At the first step in the scale development process, the literature on sustainable development was searched. As a result of the literature review, 54 items were written based on the content of the economic, environmental, and social dimensions, which are three sub-dimensions of sustainable development. Considering the opinions of four experts from different universities, the prepared item pool was re-examined by the researchers and the items that measure same features were eliminated. A final draft of 36 items was drawn up. When the sustainable development scale is divided into sub-dimensions, it is noticed that some items can be included in more than one sub-dimension. For example, item number 21 (Urbanization should be to protect the soul and body health of the society) has a characteristic that can be included in both the society and the environment sub-dimension. When deciding on to which sub-dimension such items should be included, the question of which characteristic of these items are more dominant was directed to three external experts. The item number 21 has been included in the social dimension because its social characteristic is more dominant and expert views were in this direction. Expert opinions were used to include items that correspond to regions where sub-dimensions show intersecting features. For the final form of the scale, expert views were consulted once again. The necessary revisions were made in line with the expert feedback and the scale was prepared for implementation. The draft form prepared to determine the Sustainable Development Awareness of the teacher candidates was prepared in the 5-point Likert type. In the corresponding columns of the items prepared to identify the sustainable development awareness of the teacher candidates, rating statements such as strongly disagree, disagree, neutral, agree, strongly agree were placed. The revised form of the scale was applied to the two teacher candidates at the fourth grade level in the Department of Science Education. With this application, feedback on readability and comprehensibility was provided and the

recommended time for the scale was determined. The scale was applied to 425 science teacher candidates. Confirmatory factor analysis was preferred for construct validity of the scale. Reliability analyses of the scale were performed in the SPSS package program.

### 3. RESULTS and DISCUSSION

In this section, data on the validity and reliability analyses of the scale obtained in the scale development process are presented.

#### 3.1. Validity Analysis of the Scale

Content validity of the scale was provided by expert opinion and construct validity by confirmatory factor analysis.

#### 3.2. Content Validity

For the sustainable development awareness scale, the literature was searched by the researchers and a pool of items containing 54 statements was created. The 54 items were re-examined by the researchers and the items measuring the same features were eliminated. As a result of the examination, a draft form was created with 36 items decided by the researchers. The 36 items that have been decided on include awareness about the sustainability practices regarding the three dimensions of sustainable development, namely economy, society, and environment, which are the three sub-dimensions in the literature. For the expert view, the draft form was sent to four faculty members who are in science education department of three different state universities in Turkey. In accordance with the feedback obtained from experts, the necessary revisions were made in scale items.

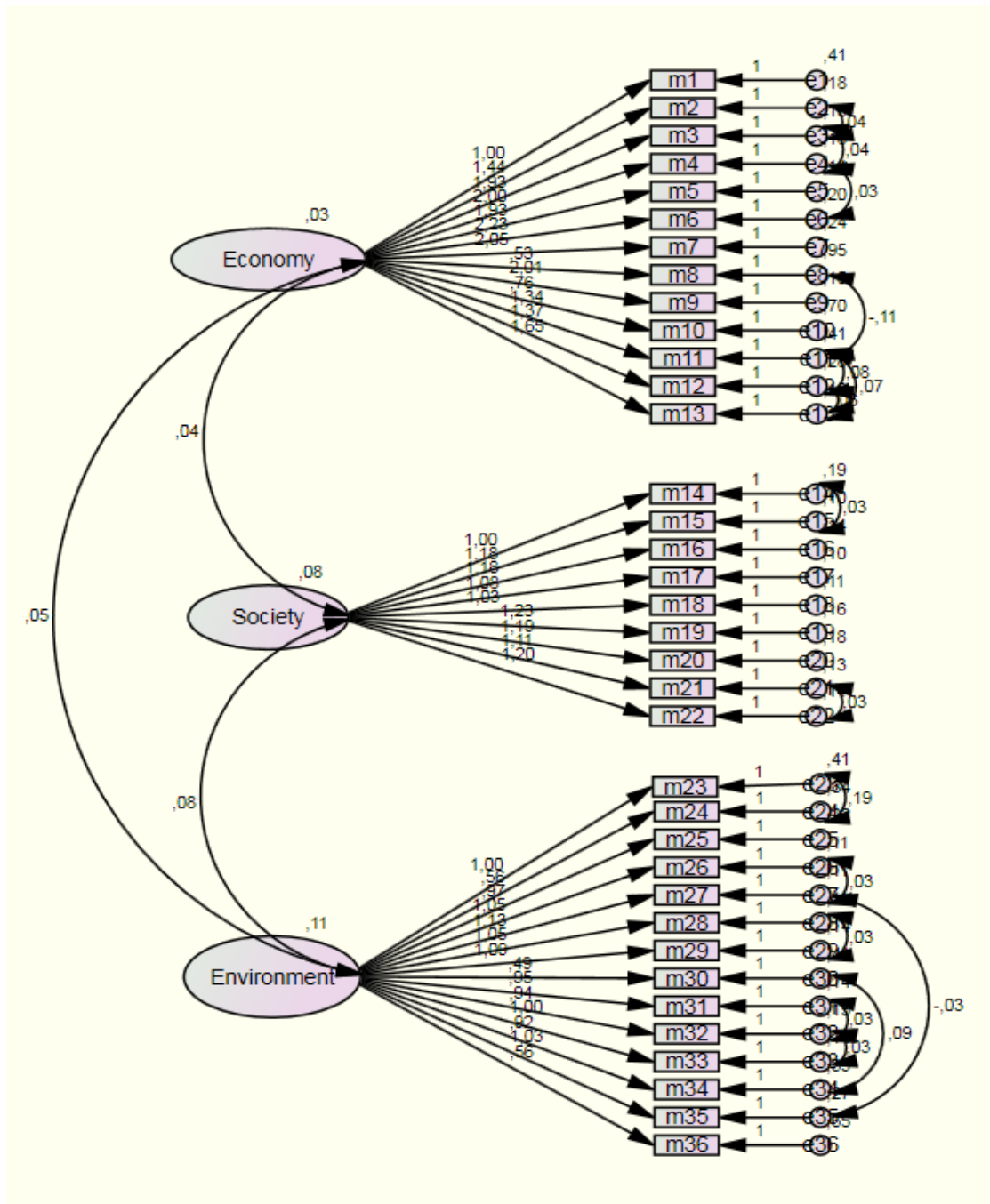
#### 3.3. Construct validity

Factor analysis was carried out in order to determine the construct validity of the scale developed for the determination of sustainable development awareness of science teacher candidates. Confirmatory factor analysis may be preferred for construct validity when the theoretical structure is evident (Akçay, Gelen, Tiryaki, & Benek, 2018; De Vellis, 2012).

In this study, construct validity was provided by confirmatory factor analysis, as scale items were prepared based on the contents of economy, environment, and society dimensions, which are three sub-dimensions of sustainable development in the literature. Confirmatory factor analysis with the aim of providing the construct validity of the Sustainable Development Awareness Scale prepared with three sub-dimensions according to theoretical framework was carried out on the AMOS program. As the multivariate normality assumption was fulfilled, the Maximum Likelihood (MLR) estimation method was used for the models.

For the confirmatory factor analysis in the study,  $\chi^2/df$  (value obtained by dividing the Chi-square fit statistic by the degree of freedom), RMSEA (Root Mean Square Error of Approximation), S-RMR (Standardized Root Mean Square Residual), AGFI (Adjusted Goodness of Fit Index), GFI (Goodness of Fit Index), IFI (Incremental Fit Index), and TLI (Trucker Lewis Index) were examined. In the confirmatory factor analysis, the fit indices used to determine whether the theoretical framework supports the data are given in [Table 1](#) (Hebecci, & Shelley, 2018; Kline, 2005; Tabachnick & Fidell, 2007). The values obtained when the analysis is repeated after the modification between the specified items as a result of the preliminary analysis were calculated as  $\chi^2 = 964.497$ ,  $df = 575$ ,  $p = .000$ ,  $\chi^2/df = 1.677$ , RMSEA = 0.40, S-RMR = 0.44, AGFI = .871, GFI = .889, IFI = .931, TLI = .923 ([Table 1](#)). When the values were evaluated on the basis of the fit indices,  $\chi^2/df$ , RMSEA, and S-RMR values were found to be in perfect fit. GFI, AGFI, IFI, and TLI values were found to be in the acceptable range. Thus, the construct of the scale is valid. The path diagram for the confirmatory factor model of the Sustainable Development Awareness Scale is shown in [Figure 1](#).





**Figure 1.** The path diagram for confirmatory factor model of Sustainable Development Awareness Scale

**Table 1.** Confirmatory Factor Model Fit Indices of Sustainable Development Awareness Scale After Modifications

| Fit Indices | Perfect Fit              | Acceptable Fit        | Fit Indices Observed in Scale Model |
|-------------|--------------------------|-----------------------|-------------------------------------|
| $\chi^2/df$ | $\chi^2/df \leq 3$       | $3 < \chi^2/df < 5$   | 1.67                                |
| RMSEA       | $0 < RMSEA \leq 0.05$    | $0.06 < RMSEA < 0.08$ | .040                                |
| S-RMR       | $0 \leq S-RMR \leq 0.05$ | $0.05 < S-RMR < 0.10$ | .044                                |
| GFI         | $GFI \geq 0.90$          | $0.85 < GFI < 0.90$   | .889                                |
| AGFI        | $AGFI \geq 0.90$         | $0.85 < AGFI < 0.90$  | .871                                |
| IFI         | $IFI \geq 0.95$          | $0.90 < IFI < 0.95$   | .931                                |
| TLI         | $TLI \geq 0.95$          | $0.90 < TLI < 0.94$   | .923                                |

### 3.4. Reliability Analysis of the Scale

The Cronbach's alpha reliability coefficient for the Sustainable Development Awareness Scale developed by the researchers was calculated for the whole scale and the sub-dimensions.

Reliability analysis of the scale was conducted with SPSS 15 package program. As a result of the analysis made, the reliability coefficient for the whole scale was calculated as Cronbach's  $\alpha = 0.91$ . The reliability coefficients for the sub-dimensions were Cronbach's  $\alpha = 0,77$  for the economy sub-dimension, Cronbach's  $\alpha = 0,87$  for the society sub-dimension, and Cronbach's  $\alpha = 0,82$  for the environment sub-dimension (Table 2).

When the reliability coefficient values calculated for the scale developed by the researchers to be used as data collection tool are examined, it is seen that a Cronbach's alpha value of .70 and above are sufficient as it is in this case (Acar, Kara, & Taşkın Ekici, 2015; Artvinli & Demir, 2018; Büyüköztürk, 2017).

**Table 2.** Reliability Coefficients of the Whole Scale and the Sub-dimensions of the Scale

| Dimensions                | Cronbach's Alpha Reliability Coefficient |
|---------------------------|--|
| Economy sub-dimension     | 0,77                                     |
| Society sub-dimension     | 0,87                                     |
| Environment sub-dimension | 0,82                                     |
| Whole Scale               | 0,91                                     |

Following the validity and reliability studies, the final form of the Sustainable Development Awareness Scale consists of 36 items with three sub-dimensions: economy, society, and environment. Items 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13 of the scale belong to the economic sustainability sub-dimension. While the items 14, 15, 16, 17, 18, 19, 20, 21, and 22 belong to the sub-dimension of social sustainability, the items 23, 24, 25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, and 37 belong to the environmental sustainability sub-dimension. Items 1, 8, 10, 24, 31, and 35 are negative and the item 26 is control. The lowest score to be taken from the scale was calculated as 36 and the highest score as 180 (Appendix 1).

### 4. CONCLUSION

It is of great importance that the sustainable development awareness occurs in countries where industrialization and urbanization continue rapidly. With this scale developed, it is aimed to determine the awareness of science teacher candidates, who are responsible for ensuring sustainable development awareness in society, before they begin to the profession. For this reason, this study has been carried out with the aim of developing a valid and reliable measurement tool to determine the awareness of science teacher candidates about sustainable development. The Sustainable Development Awareness Scale developed in this study has three sub-dimensions as economy, society, and environment in accordance with the theoretical framework accepted in the literature. The sub-dimensions of the 36-item Sustainable Development Awareness Scale were supported using confirmatory factor analysis. Findings obtained after the analyses support the fact that the developed scale is a suitable measurement tool for determining the sustainable development awareness of science teacher candidates.

The fact that item number of the scale developed in this study is more compared to the previous scales increased the content validity of the scale. There are 14 items that measure the environmental sub-dimension of the scale, 13 items that measure the economy sub-dimension, and 9 items that measure the social sub-dimension. These items reflect the features of these sub-dimensions in a more detailed and comprehensive way than previously developed scales.

While this scale is being developed, the economic, social, and environmental sub-dimensions were designed to be independent of each other. When the items fell into the intersection areas of these three dimensions, it has been decided by experts which dimension the item should be included. As a recommendation to the researchers who want to develop a sustainable development awareness scale in the next stage, a seven sub-dimensional scale may be proposed, which considers the intersection areas of sub-dimensions as separate sub-dimensions. This scale was developed for science teacher candidates. However, it is expected that they will give similar results when working with science teachers. It may also be recommended to carry out studies of reliability and validity by applying the scale to science teachers.

### Acknowledgment

The English version of this scale is in the end of this article. However, the English version is not suitable to use because its language validity has not been analyzed. The English version of the scale is just for the readers' curiosity.

This study has been derived from the first author's master thesis.

### ORCID

Seyit Ahmet Kiray  <https://orcid.org/0000-0002-5736-2331>

### 5. REFERENCES

- Acar, C., Kara, I., & Taşkın Ekici, F. (2015). Development of Self Directed Learning Skills Scale for Pre-Service Science Teachers. *International Journal of Assessment Tools in Education*, 2(2), 3-13.
- Akçay, B., Gelen, B., Tiryaki, A., & Benek, I. (2018). An analysis of scale adaptation studies in science education: Meta-synthesis study. *Journal of Education in Science, Environment and Health (JESEH)*, 4(2), 227-245. DOI:10.21891/jeseh.439150
- Alkış, S. (2007). Coğrafya eğitiminde yükselen paradigma: Sürdürülebilir bir dünya. [The rising paradigm in teaching geography: a sustainable world] *Marmara Coğrafya Dergisi*, 15, 55-64.
- Altunbaş, D. (2003). Uluslararası sürdürülebilir kalkınma ekseninde Türkiye'deki kurumsal değişimlere bir bakış. [A new perspective towards the institutional changes in Turkey within the frame of international sustainable development] *Yönetim Bilimleri Dergisi*, 1(1-2), 103-118.
- Artvinli, E., & Demir, Z. M. (2018). A study of developing an environmental attitude scale for primary school students. *Journal of Education in Science, Environment and Health (JESEH)*, 4(1), 32-45. DOI:10.21891/jeseh.387478
- Atmaca, A.C., Kiray, S.A., & Pehlivan, M. (2018). Sustainable Development from Past to Present. In Shelley, M. & Kiray, S.A.(Ed.). *Education Research Highlights in Mathematics, Science and Technology 2018* (pp. 186-214). ISRES Publishing, ISBN: 978-605-81654-3-4. <https://www.isres.org/education-research-highlights-in-mathematics-science-and-technology-2018-6-b.html#.XCPdZ1wzZPY>
- Aydoğan, A. (2010). Sosyal bilgiler öğretmenlerinin sürdürülebilir kalkınma konusuyla ilgili kazanımların öğretimine ilişkin görüşleri. [Social studies teachers' views on the teaching of achievements related to sustainable development] Yüksek Lisans Tezi. Niğde Üniversitesi Sosyal Bilimler Enstitüsü İlköğretim Anabilim Dalı, Niğde.
- Baykal, H., & Baykal, T. (2008). Küreselleşen dünyada çevre sorunları. [Environmental problems in a globalized world] *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 5(9).

- Biasutti, M., & Surian, A. (2012). The students' survey of education for sustainable development competencies: A comparison among faculties. *Discourse and Communion for Sustainable Development*, 3(1), 75-82.
- Biasutti, M., & Frate, S. (2017). A validity and reliability study of the attitudes toward sustainable development scale. *Environmental Education Research*, 23(2), 214-230.
- Borg, C., Gericke, N., Höglund, H. O., & Bergman, E. (2012). The barriers encountered by teachers implementing education for sustainable development: Discipline bound differences and teaching traditions. *Research in Science & Technological Education*, 30(2), 185-207.
- Büyüköztürk, Ş. (2017). *Sosyal bilimler için veri analizi el kitabı [The data analysis manual for the social sciences]* (23. Baskı). Ankara. Pegem Akademi.
- Chow, W. S., & Chen, Y. (2012). Corporate sustainable development: Testing a new scale based on the Mainland Chinese context. *Journal of Business Ethic*, 105(4), 519-533.
- Demirbaş, Ç. Ö. (2015). Öğretmen adaylarının sürdürülebilir kalkınma farkındalık düzeyleri. [Sustainable development awareness levels of teacher candidates.] *Marmara Coğrafya Dergisi*, 31, 300-3016.
- DeVellis, R.F. (2012). *Scale development: Theory and applications*. Sage Publications, Thousand Oaks, CA.
- Doğan, O., Bulut, Z. A., & Çımrın, F. K. (2015). Bireylerin sürdürülebilir tüketim davranışlarının ölçülmesine yönelik bir ölçek geliştirme çalışması. [A scale development study to measure individuals' sustainable consumption behavior] *İktisadi ve İdari Bilimler Dergisi*, 29(4), 659-678.
- Erten, S. (2015). Sample course material for biodiversity and sustainable education. *International Journal of Education in Mathematics, Science and Technology*, 3(2), 155-161.
- Hebecci, M.T., & Shelley, M. (2018). *International Journal of Assessment Tools in Education*, 5(2), 223-234.
- Kahyaoğlu, M. (2011). Öğretmen adaylarının öğrenme stilleri ile çevre eğitimi öz-yeterlilikleri arasındaki ilişki. [Relationship between the Self Efficacy Beliefs towards Environmental Education and the Learning Styles of Pre-service Teachers] *Eğitim Bilimleri Araştırmaları Dergisi*, 1(2), 68-82.
- Kaya, M. F. (2013). Sürdürülebilir kalkınmaya yönelik tutum ölçeği geliştirme çalışması. [A Scale development study on the attitudes of sustainable development] *Marmara Coğrafya Dergisi*, 28, 175-193.
- Koçak, F., & Balcı, V. (2010). Doğada yapılan sportif etkinliklerde çevresel sürdürülebilirlik. [The environmental sustainability in the sporting events in nature ] *Ankara Üniversitesi Çevre Bilimleri Dergisi*, 2(2), 213-222.
- Kuşat, N. (2013). Yeşil sürdürülebilirlik için yeşil ekonomi: Avantaj ve dezavantajları Türkiye incelemesi. [Green economy for green sustainability: advantages and disadvantages – turkey review] *Journal of Yasar University*, 29(8), 4896-4916.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.
- Manju, N. D. (2015). A study on opinion of teachers about role of education in sustainable development. *International Journal of Education and Psychological Research*, 4(1), 60-64.
- Olsson, D., Gericke, N., & Chang Rundgren, S. N. (2016). The effect of implementation of education for sustainable development in Swedish compulsory schools: Assessing pupils' sustainability consciousness. *Environmental Education Research*, 22(2), 176-202.

- Özmete, E., & Akgül Gök, F. (2015). Sürdürülebilir kalkınma için sosyal inovasyon ve sosyal hizmet ilişkisinin değerlendirilmesi.[Evaluation of relationship between social innovation and social work for sustainable development] *Hacettepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi Sosyal Hizmet Bölümü Dergisi*, 26(2), 127-143.
- Sandel, K., Öhman, J., & Östman, L. (2006). *Education for sustainable development: Nature, school and democracy*. Lund: Student litteratur.
- Summers, M., Kruger, C., Childs, A., & Mant, J. (2010). Primary school teachers' understanding of environmental issues: An interview study. *Environmental Education Research*, 6(4), 294-312.
- Şahin, İ., & Kutlu, S. Z. (2014). Cittaslow: Sürdürülebilir kalkınma ekseninde bir değerlendirme. [Cittaslow: an assessment from the perspective of sustainable development] *Journal of Tourism and Gastronomy Studies*, 2(1), 55-63.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental designs using ANOVA*. Thomson/Brooks/Cole
- Türer, B. (2010). *Fen bilgisi ve sosyal bilgiler öğretmen adaylarının sürdürülebilir kalkınma farkındalıklarının belirlenmesi.[The awareness levels of science and social science prospective teachers regarding sustainable environment]* Yüksek Lisans Tezi. Ondokuz Mayıs Üniversitesi, Eğitim Bilimleri Enstitüsü, İlköğretim Anabilim Dalı, Samsun.
- United Nations. (1987). *Report of the world commission on environment and development: Our common future*. <http://www.un-documents.net/our-common-future.pdf> (Reached on 02.12.2017).
- UNESCO. (2006). *United Nations decade of education for sustainable development 2005–2014, UNESCO: International implementation scheme*. Paris: UNESCO.
- WCED. (1987). *Our common future: A report from the United Nations World Commission on Environment and Development*. Oxford: Oxford University Press.
- Yerdelen, S., Cansız, M., Cansız, N., & Akcay, H. (2018). Promoting preservice teachers' attitudes toward socioscientific issues. *Journal of Education in Science, Environment and Health (JESEH)*, 4(1), 1- 11. DOI:10.21891/jeseh.387465



**Appendix 1.** Sustainable Development Awareness Scale Items (English version)

| <i>Items</i> |   |
|--------------|---|
| 1            | Individuals should shop in the direction of their desires and wishes without regard to their needs.   |
| 2            | We must use current economic resources with conservation, thinking about future generations.  |
| 3            | Debt to be made for development should be made considering economic balances.   |
| 4            | Economic policies should be able to reduce poverty and differences in income distribution.  |
| 5            | Economic development should be planned to prevent unemployment.   |
| 6            | Economic policies should be shaped by sustainable production.   |
| 7            | Economic policies should be shaped so as not to destroy natural resources.  |
| 8            | Livestock, agricultural and industrial production should be focused on applications that will generate high profits in the short term (use of GMO products, hormonal animals etc.). |
| 9            | For economic investments, environments where life and property safety are provided must be established.   |
| 10           | For economic development, non-production sectors should be emphasized.  |
| 11           | The production of high-tech products for economic development should be supported.  |
| 12           | Investments in agriculture and livestock sectors should be supported for economic development.  |
| 13           | Research and development (R&D) studies for economic development should be supported.  |
| 14           | Equal opportunities should be offered to individuals in society (women/men, rich/poor, race/religion etc.).   |
| 15           | For all individuals in society, environments should be created to enable the individual to learn life-long.   |
| 16           | Individuals should be provided with integrating and enhancing social services (such as nurseries, shelter homes, social assistance foundations etc.).                               |
| 17           | Access to education and health services should be provided to all individuals in society.   |
| 18           | Individuals should be provided with environments where they feel safe while living.   |
| 19           | Interaction of cultures in society should be supported and developed.   |
| 20           | The society must take responsibility to keep the well-being of individuals and families above the minimum.  |
| 21           | Urbanization (city, town, etc.) should be to protect the soul and body health of the society.   |
| 22           | The work of governmental and non-governmental organizations involved in activities for the sustainable environment should be supported.   |
| 23           | Any intervention that damages natural life (wrong use of pesticide, prohibited hunting, etc.) must be punished for the continuation of biological diversity.                        |
| 24           | The use of public transportation at short distances does not help to maintain atmospheric equilibrium.  |
| 25           | I think that vehicles with the least impact on degradation of ecological balance should be preferred when buying one.   |
| 26           | Energy saving products should be preferred in order to use energy sources for a longer time.  |
| 27           | The use of renewable energy sources needs to be widespread to leave a livable world.  |
| 28           | Every individual has responsibility to protect existing resources (water, air, soil etc.) for future generations to survive ecological problems.                                    |
| 29           | Industrial establishments should take cautions to protect environmental health and prevent pollution of natural resources.  |
| 30           | Green areas can be dispensed with for urbanization and industrialization.   |
| 31           | In order to leave a greener world for future generations, responsibility for afforestation and the protection of the trees is the responsibility of each individual.                |
| 32           | I think that each individual has responsibilities in the process of recycling wastes so that the raw material resources can be used by future generations.                          |
| 33           | Wastes should be separated according to their characteristics and reused, so that raw material sources can be used by future generations.   |
| 34           | I think that nothing can be done individually to prevent global climate change.   |
| 35           | I think global warming poses a serious threat to the future of our world if cautions are not taken.   |
| 36           | I think that ecological footprint should be minimized for the continuation of the world's livability.   |

**Appendix 2.** Sustainable Development Awareness Scale Items (Turkish version)

| <i>Maddeler</i> |  |
|-----------------|--|
| 1               | Bireyler ihtiyaçlarını gözetmeksizin, arzu ve istekleri doğrultusunda alışveriş yapmalıdır.  |
| 2               | Gelecek nesilleri de düşünerek mevcut ekonomik kaynakları tasarruflu kullanmalıyız.  |
| 3               | Kalkınma için yapılacak borçlanma ekonomik dengeler gözetilerek yapılmalıdır.  |
| 4               | Ekonomik politikalar, yoksulluğu ve gelir dağılımındaki farklılıkları azaltıcı nitelikte olmalıdır.  |
| 5               | Ekonomik kalkınma işsizliği önleyecek şekilde planlanmalıdır.  |
| 6               | Ekonomik politikalar sürdürülebilir üretime göre şekillenmelidir.  |
| 7               | Ekonomik politikalar doğal kaynakları yok etmeyecek şekilde oluşturulmalıdır.  |
| 8               | Hayvancılık da, tarımsal ve endüstriyel üretim de, kısa vadede yüksek kâr elde edecek uygulamalara (GDO' lu ürün kullanımı, hormonlu hayvanlar v.b.) ağırlık verilmelidir. |
| 9               | Ekonomik yatırımlar için can ve mal güvenliğinin sağlandığı ortamlar oluşturulmalıdır.   |
| 10              | Ekonomik kalkınma için üretim dışı sektörler ağırlık verilmelidir.   |
| 11              | Ekonomik kalkınma için ileri teknoloji ürünlerinin üretimi desteklenmelidir.   |
| 12              | Ekonomik kalkınma için tarım ve hayvancılık sektörlerine yapılacak yatırımlar desteklenmelidir.  |
| 13              | Ekonomik kalkınma için araştırma geliştirme (AR-GE) çalışmaları desteklenmelidir.  |
| 14              | Toplumdaki bireylere (kadın/erkek, zengin/fakir, ırk/din v.b.) eşit fırsatlar sunulmalıdır.  |
| 15              | Toplumdaki bütün bireyler için bireyin yaşam boyu öğrenmesine olanak sağlayacak ortamlar oluşturulmalıdır.   |
| 16              | Bireylere, toplumla bütünleştirici ve geliştirici sosyal hizmetler (çocuk yuvaları, huzur evi, sosyal yardımlaşma vakıfları v.b.) sunulmalıdır.                            |
| 17              | Toplumdaki bütün bireylere eğitim ve sağlık hizmetlerine ulaşım hakkı sağlanmalıdır.   |
| 18              | Bireylere, yaşadıkları yerlerde kendilerini güvende hissedebilecekleri bir ortam oluşturulmalıdır.   |
| 19              | Toplumda ki kültürlerin birbiri ile etkileşimi desteklenmeli ve geliştirilmelidir.   |
| 20              | Bireylerin ve ailelerin refah düzeyini asgari koşulların üzerinde tutmak için toplum sorumluluk almalıdır.   |
| 21              | Şehirleşme, (şehir, kasaba v.b) toplumun ruh ve beden sağlığını koruyacak şekilde olmalıdır.   |
| 22              | Sürdürülebilir çevre için faaliyetlerde bulunan resmi ve sivil toplum kuruluşlarının çalışmaları desteklenmelidir.   |
| 23              | Biyolojik çeşitliliğin devamı için doğal yaşama zarar veren her müdahale (bilinçsiz ilaçlama, yasak avlanma vb.) cezalandırılmalıdır.                                      |
| 24              | Kısa mesafelerde toplu taşıma araçları kullanılmasının atmosferik dengenin korunmasına faydası yoktur.   |
| 25              | Araç alırken, ekolojik dengenin bozulmasına etkisi en az olan araçların tercih edilmesi gerektiğini düşünüyorum.   |
| 26              | Enerji kaynaklarının daha uzun süreli kullanılabilmesi için enerji tasarrufu yapan ürünlerin tercih edilmesi gerekir.  |
| 27              | Yaşanılabilir bir dünya bırakabilmek için yenilenebilir enerji kaynaklarının kullanımının yaygınlaştırılması gerekir.  |
| 28              | Gelecek nesillerin ekolojik sorunlar yaşamaması için mevcut kaynakların (su, hava, toprak v.b.) korunması hususunda her bireye düşen sorumluluklar vardır.                 |
| 29              | Endüstri kuruluşları çevre sağlığını koruyacak ve doğal kaynakların kirletilmesini önleyecek tedbirler almalıdır.  |
| 30              | Yeşil alanlardan şehirleşme ve sanayileşme amacıyla vazgeçilebilir.  |
| 31              | Gelecek nesillere daha yeşil bir dünya bırakabilmek için ağaçlandırma çalışmaları ve ağaçların korunması ile ilgili her bireye sorumluluk düşmektedir.                     |
| 32              | Ham madde kaynaklarının gelecek nesiller tarafından da kullanılabilmesi için atıkların geri dönüştürülmesi sürecinde her bireyin sorumlulukları olduğunu düşünüyorum.      |
| 33              | Ham madde kaynaklarının gelecek nesiller tarafından da kullanılabilmesi için çöpler özelliklerine göre ayrılarak, değerlendirilmelidir.                                    |
| 34              | Küresel iklim değişikliğini önlemek için bireysel olarak hiçbir şey yapılamayacağını düşünüyorum.  |
| 35              | Önlem alınmaması halinde küresel ısınmanın, dünyamızın geleceği için ciddi tehdit oluşturduğunu düşünüyorum.   |
| 36              | Dünyanın yaşanabilirliğinin devamı için ekolojik ayak izimizin küçültülmesi gerektiğini düşünüyorum.   |

## The Impact of Ignoring Multilevel Data Structure on the Estimation of Dichotomous Item Response Theory Models

Hyung Rock Lee <sup>1,\*</sup>, Sunbok Lee <sup>2,</sup> Jaeyun Sung <sup>3</sup>

<sup>1</sup> University of Central Arkansas, Department of Exercise & Sport Science, Conway, AR USA

<sup>2</sup> University of Houston, Department of Psychology, Houston, TX USA

<sup>3</sup> Lyon College, Department of Political Science, Batesville, AR, USA

### ARTICLE HISTORY

*Received: 14 November 2018*

*Accepted: 05 February 2019*

### KEYWORDS

Item Response Theory,  
Rasch,  
Multilevel Data,  
Monte Carlo Simulation

**Abstract:** Applying single-level statistical models to multilevel data typically produces underestimated standard errors, which may result in misleading conclusions. This study examined the impact of ignoring multilevel data structure on the estimation of item parameters and their standard errors of the Rasch, two-, and three-parameter logistic models in item response theory (IRT) to demonstrate the degree of such underestimation in IRT. Also, the Lord's chi-square test using the underestimated standard errors was used to test differential item functioning (DIF) to show the impact of such underestimation on the practical applications of IRT. The results of simulation studies showed that, in the most severe case of multilevel data, the standard error estimate from the standard single-level IRT models was about half of the minimal asymptotic standard error, and the type I error rate of the Lord's chi-square test was inflated up to .35. The results of this study suggest that standard single-level IRT models may seriously mislead our conclusions in the presence of multilevel data, and therefore multilevel IRT models need to be considered as alternatives.

## 1. INTRODUCTION

In traditional statistical models, observations are typically assumed to be independent. However, the assumption of independence is quite strong and may not be tenable in practice. In educational research, for example, observations in data are often not independent because of a hierarchical data structure. It is well known that applying traditional statistical models based on the independence assumption to multilevel data may result in incorrect standard errors (Barcikowski, 1981; Tate & Wongbundhit, 1983; Satorra & Muthen, 1995; Goldstein, 1987; Julian, 2001; Finch & French, 2011). Because the use of correct standard errors is the key

---

**CONTACT:** Hyung Rock Lee ✉ [rlee@uca.edu](mailto:rlee@uca.edu) 📧 University of Central Arkansas, Department of Exercise & Sport Science, Conway, AR USA

ISSN-e: 2148-7456 /© IJATE 2019

element for valid statistical inferences such as hypothesis testing and confidence intervals, applying single-level models to multilevel data could be problematic.

Given the concerns on the use of single-level models to multilevel data, the goal of this study is to examine the extent to which multilevel data structure affects the estimation of the single-level dichotomous IRT models and their subsequent application. More specifically, two Monte Carlo simulation studies were conducted to examine 1) the impact of ignoring multilevel data structure on the estimation of item parameters and their standard errors of the standard single-level Rasch, two- (2PL), and three- (3PL) parameter logistic models in item response theory (IRT); 2) the type I error inflation of the Lord's chi-square tests based on standard errors estimated from the single-level IRT models. In the simulation study 1, item responses with multilevel data structure were generated using the Rasch, 2PL, and 3PL models formulated in the hierarchical generalized linear model (HGLM), in which items, persons, and schools were modeled in Level-1, Level-2, and Level-3, respectively (Kamata & Vaughn, 2011). In generating item responses with multilevel structure, intraclass correlation coefficients (ICCs), numbers of groups, and group sizes were manipulated. Given the item responses with multilevel structure, item parameters and their standard errors were estimated using single-level IRT models with BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). To evaluate the extent to which standard errors are underestimated, the analytical minimal standard errors (Thissen & Wainer, 1982) for item parameters in the Rasch, 2PL, and 3PL were used as reference values. In practice, the underestimated standard errors can be used for other applications such as DIF tests. In the simulation study 2, the type I error rates of two DIF tests were compared: the Lord's chi-square test (Lord, 1980) using the underestimated standard errors from the single-level IRT models and the DIF test based on the Rasch model that was formulated in the hierarchical generalized linear model (HGLM).

## **2. IGNORING MULTILEVEL DATA STRUCTURE IN STATISTICAL MODELS**

The impact of multilevel data on the estimation of standard errors in statistical models can be illustrated by an example of cluster sampling designs (Kish, 1965), in which only a subset of primary units or clusters is randomly selected, and then secondary units are sampled within the selected primary units. Cluster sampling designs are often preferred because of cost and time effectiveness. In cluster sampling designs, respondents in the same cluster are likely to be similar to one another because they share similar contexts. From a statistical viewpoint, the similarity between respondents makes the information in data more redundant or less unique, which results in the reduction of effective sample sizes. As a result, the estimated sampling variances from cluster sampling designs are larger than the ones from simple random sampling designs. The loss of effectiveness in cluster sampling designs is measured by the design effect, which is defined as the ratio of the sampling variance in cluster sampling designs to the sampling variance in simple random sampling designs. In other words, the design effect is a correction factor to be multiplied to the sampling variance of the simple random sampling to get the actual sampling variance in cluster sampling designs (Hox, 1998). In the simplest cluster sampling design, the design effect is defined by the following equation:

$$\text{Design effect} = 1 + (n_g - 1)\rho_I, \quad (1)$$

where  $n_g$  is the sample size in a group, and  $\rho_I$  is the intraclass correlation coefficient (ICC). The ICC provides a measure of the amount of dependency among individuals or how similar individuals are within groups. As can be seen from Equation 1, the design effect is greater than one for a non-zero ICC. Therefore, if appropriate statistical models that can accommodate cluster structure are used, the sampling variance in data from cluster sampling designs should be larger than the one from simple random sampling designs because of reduction in effective

sample sizes. On the other hand, given non-zero ICCs, observed variance within clusters is typically less than observed variance between clusters since observations within cluster tend to be more similar to one another. Therefore, when observations are assumed to be independent, overall observed variance that is obtained without reflecting cluster nature tends to be underestimated, which could result in type I error inflation (Goldstein, 1987).

The effect of multilevel data structure on the estimation of statistical models has been investigated in many different settings. Barcikowski (1981) reported that the type I error rates of t-tests can be dramatically increased as the ICC increased. Also, Tate and Wongbundhit (1983) reported that the ordinary least square (OLS) regression produced unbiased parameter estimates but downwardly biased standard error estimates in the presence of multilevel data. Satorra and Muthen (1995) compared the standard maximum likelihood estimation, the robust maximum likelihood estimation, and the multilevel maximum likelihood estimation for structural equation modeling (SEM) under complex sampling designs and found that standard error estimates from the standard maximum likelihood estimation were downwardly biased. Recently, Finch and French (2011) found that applying standard approaches for differential item functioning (DIF) to multilevel data caused type I error inflation. In line with those concerns on the use of standard single-level models in the presence of multilevel data, this study was designed to explicitly show the degree to which the multilevel data structure influences the estimation of standard single-level IRT models.

### 3. STANDARD ERRORS IN IRT APPLICATIONS

Standard errors measure the accuracy of estimation. Using correct standard errors is an essential component for valid statistical inferences based on hypothesis tests and confidence intervals. The use of the correct standard error is also important in many IRT applications (Toland, 2008). For example, the accurate standard error estimate is important in identifying DIF items using the Lord's chi-square test in which the difference in the item parameters between the focal and reference groups is tested using the following equation:

$$\chi^2 = \frac{(\hat{\theta}_F - \hat{\theta}_R)^2}{\hat{\sigma}_F^2 + \hat{\sigma}_R^2}, \quad (2)$$

Where  $\hat{\theta}_F$  and  $\hat{\theta}_R$  represent the parameter estimates in the focal and reference groups, and  $\hat{\sigma}_F^2$  and  $\hat{\sigma}_R^2$  represent the standard error estimates for  $\hat{\theta}_F$  and  $\hat{\theta}_R$ . Because the item parameter estimates for the focal and reference groups are obtained from separate calibrations, item parameter estimates need to be transformed on a common metric using an appropriate transformation. As can be seen from Equation 2, standard error estimates affect the result of the Lord's chi-square test. Some other IRT applications also require accurate standard error estimates for item parameter estimates (Toland, 2008): the separate calibration *t*-test for DIF (Wright & Stone, 1979), the item parameter replication (IPR) method for testing non-compensatory DIF (Oshima, Raju, & Nanda, 2006), and the cumulative sum (CUSUM) procedure for the computer adaptive test (Veerkamp & Glas, 2000).

In examining the estimation for standard errors in IRT models, this study used the minimum obtainable standard errors for item parameters (Thissen & Wainer, 1982) as references values. Thissen and Wainer (1982) derived analytical asymptotic standard errors for item parameters using the inverse information matrix. Those asymptotic standard errors can be considered as the lower limits for the estimated standard errors because they are derived under the very strong assumptions which are not likely to be met in practice. Therefore, estimated standard errors are larger than the minimal asymptotic standard errors.



#### 4. MULTILEVEL IRT MODELS

One of the assumptions in traditional IRT models is the local independence assumption in which the dependencies among item responses are assumed to be fully explained by the specified IRT model (Embretson & Reise, 2000). More specifically, two different kinds of local independence assumptions can be considered (Reckase, 2009; Jiao, Kamata, Wang, & Jin, 2012). The local item independence refers to the independence of responses for items within a specific person. Given the ability of a person, a person's response to an item does not have any influence on the probability of that person's response to another item. On the other hand, the local person independence refers to the independence of responses of persons for a specific item. Given the abilities of persons, a person's response to a specific item does not affect the probability of another person's response to that item.

Since the traditional IRT models assume a single source of the dependencies among item responses, which is the ability of a person, problems could occur when the dependencies among item responses still remain beyond what is explained by the specified IRT model. In order to fully explain the dependencies, therefore, additional sources of the dependencies need to be specified in the IRT model. For example, a common passage in a test could cause additional dependencies among item responses. In that case, the local item independence is considered to be violated. On the other hand, the local person independence could be violated in the presence of clustered data (Jiao et al., 2012). For example, the responses of students from the same school could be more similar to each other than to responses from students from other schools, even after controlling for the abilities of persons. In multilevel IRT models, the clustered data structure is considered the additional source of the dependencies among item responses (Kamata, 2001).

A simple multilevel IRT model assumes that items are nested within persons, and persons are nested within groups (Kamata & Vaughn, 2011). For example, multilevel 2PL models can be expressed as

$$P_{ijg}[Y = 1] = \frac{\exp[\alpha_i(\theta_g + \theta_{jg}) + \beta_i]}{1 + \exp[\alpha_i(\theta_g + \theta_{jg}) + \beta_i]} \quad (3)$$

Where  $\alpha_i$  and  $\beta_i$  are the discrimination and difficulty parameters of item  $i$ ,  $\theta_g$  is the mean of ability of group  $g$ ,  $\theta_{jg}$  is the amount of deviation from the group mean ability for a person  $j$  in a group  $g$ .

#### 5. SIMULATION STUDY1

##### 5.1. Simulation Designs

This simulation study was designed to examine the impact of ignoring multilevel data structure on the estimation of the Rasch, 2PL, and 3PL models. To simulate multilevel data structure, item responses were generated based on Equation 4 below (Kamata & Vaughn, 2011). The parameters and their standard errors were estimated using BILOG-MG (Zimowski et al., 1996). To make estimates comparable across replications, metric transformations were performed to put estimates on a common scale. This simulation was conducted using the R software package (R Core Team, 2013).

##### 5.1.1. Simulation Conditions

The simulation conditions for multilevel data structure was manipulated in terms of the ICC, number of groups (nG), and group sizes (nW). The values of the ICC in this simulation study were set at 0, .05, .15, .25, .35, and .45 based on prior research. Hedges and Hedberg (2007) reported that the values of the ICC in educational performance

research often range between .10 and .25. Snijders and Bosker (1999) reported that the values of the ICC between .05 and .20 are most common in educational research, and values greater than .20 can be considered large. Also, the numbers of groups were set at 50, 100, and 200 based on prior research (Maas & Hox, 2005; Finch & French, 2011). The group sizes or within-group sample sizes were set at 5, 15, 25, and 50, which cover the typical range of within-group sample sizes in family and educational research (Maas & Hox, 2005). In all, there were total 72 ( $=6 \times 3 \times 4$ ) simulation conditions, and 1000 simulated data sets were replicated for each simulation condition.

### 5.1.2. Data Generation

To simulate multilevel data structure, item responses were generated based on the following equation:

$$P_{ijg}[Y = 1] = r_i + (1 - r_i) \frac{\exp[\alpha_i(\theta_g + \theta_{jg}) + \beta_i]}{1 + \exp[\alpha_i(\theta_g + \theta_{jg}) + \beta_i]} \quad (4)$$

$$\theta_{jg} \sim N(0,1), \quad (5)$$

$$\theta_g \sim N\left(0, \sigma_{\theta_g}^2\right), \quad (6)$$

which is the three-level hierarchical generalized linear model (Kamata, 2001), in which items, persons, and groups are modeled in Level-1, Level-2, and Level-3, respectively. In this simulation, the values of the difficulty parameters for seven items were set at (-3, -2, -1, 0, 1, 2, 3) so that the estimated standard errors can be compared to the minimal asymptotic standard errors tabulated in Thissen and Wainer (1982). The values of the discrimination parameters of the Rasch, 2PL, and 3PL models were set at (1, 1, 1, 1, 1, 1, 1), (1, 2, 1, 2, 1, 2, 1), and (1, 2, 1, 2, 1, 2, 1), respectively. For the 3PL model, guessing parameters were set at (0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2).

The proportion of the between-group variance in the total variance, which is the ICC, was calculated based on the following equation:

$$ICC = \frac{\theta_{\theta_g}^2}{\theta_e^2 + \sigma_{\theta_{jg}}^2 + \sigma_{\theta_g}^2}, \quad (7)$$

Where  $\sigma_e^2$ ,  $\sigma_{\theta_{jg}}^2$ , and  $\sigma_{\theta_g}^2$  represent the variation in Level-1, Level-2, and Level-3, respectively.  $\sigma_e^2$  representing the Level-1 variance was set at  $\pi^2/3$  following Snijders and Bosker (2011).  $\sigma_{\theta_{jg}}^2$  representing the amount of deviation from the group mean ability for a person  $j$  in a group  $g$  and was set at 1. The values of  $\sigma_{\theta_g}^2$ , which represents the variance of the mean of abilities in a group  $g$ , can be determined given the values of the ICC.

### 5.1.3. Minimal Asymptotic Standard Errors

In examining the estimated standard errors from BILOG-MG, the minimal asymptotic standard errors (Thissen & Wainer, 1982) were used as reference values. Thissen and Wainer (1982) provided tables that contain minimal asymptotic standard errors for various values of locations, slopes, and asymptote parameters. Note that the values in the tables need to be adjusted using specific values of sample sizes.

### 5.1.4. Scale Transformation

In estimating parameters in IRT models, some parameters need to be fixed to arbitrary values to identify the models. Therefore, in IRT, independent estimates from two separate data sets can be compared only after they are expressed on a common metric (Stocking & Lord, 1983).

In this study, item parameter estimates from each replication were transformed into the metric defined by the original parameter values using the following equations (De Ayala, 2009):  $\hat{\alpha}^* = \hat{\alpha}/A$ ,  $\hat{\beta}^* = A\hat{\beta} + B$ ,  $\hat{c}^* = \hat{c}$ , where  $A = S_{\hat{\beta}^*}/S_{\hat{\beta}}$ , and  $B = \bar{\beta}^* - A\bar{\beta}$ ;  $\hat{\alpha}$ ,  $\hat{\beta}$ , and  $\hat{c}$  represent the discrimination, difficulty, and guessing parameter estimates in the original metric;  $\hat{\alpha}^*$ ,  $\hat{\beta}^*$ , and  $\hat{c}^*$  represent corresponding estimates in the target metric; and  $S_{\hat{\beta}^*}$  and  $S_{\hat{\beta}}$  represent standard deviations for difficulty parameters on the target and original metric respectively. The standard error estimates were also transformed to the metric defined by the original parameter values using the following equations (Kim & Cohen, 1995):

$$SE = \sqrt{Var[\hat{\alpha}^*]} = \sqrt{Var\left[\frac{\hat{\alpha}}{A}\right]} = \frac{SE[\hat{\alpha}]}{A}, \quad (8)$$

$$SE[\hat{\beta}^*] = \sqrt{Var[\hat{\beta}^*]} = \sqrt{Var[A\hat{\beta} + B]} = A \times SE[\hat{\beta}], \quad (9)$$

Where the coefficients A and B are the ones that are defined above.

### 5.1.5. Evaluation

To evaluate the impact of multilevel data structure on the estimation of item parameters for standard IRT models, the bias was calculated using the following equation and compared across simulation conditions:

$$Bias(\hat{\theta}) = \frac{\sum_{r=1}^R \sum_{i=1}^I (\hat{\theta}_{ri} - \theta_i)}{RI}, \quad (10)$$

Where  $R$  and  $I$  represent the number of replications and the number of items respectively. Also, the following ratio was calculated to compare the standard error estimates from BILOG-MG with the minimal asymptotic standard errors (Thissen & Wainer, 1982):

$$r = \frac{SE_B}{SE_T}, \quad (11)$$

Where  $SE_B$  and  $SE_T$  represent the standard error estimates from BILOG-MG and the minimal asymptotic standard errors, respectively.

On the other hand, the type I error inflation is also of interest when the standard errors are underestimated. To obtain a rough idea for the type I error inflation in the presence of underestimated standard errors, the theoretical type I errors of the  $z$ -tests for the statistical significance of item parameters were calculated in the following way. Under the assumption that item parameters following the standard normal distribution, the type I errors can be expressed as the following:

$$Type\ I\ error = 1 - \int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz, \quad (12)$$

Now, let us express the  $z$ -statistic based on the standard error estimates from BILOG-MG, which is denoted by  $z'$ , in terms of the  $z$ -statistics based on the minimal asymptotic standard error estimates, which is denoted by  $z$ , as the following:

$$z' = \frac{\theta}{SE_B} = \frac{\theta}{rSE_T} = \frac{z}{r}, \quad (13)$$

Under the assumption that the  $z$ -test based on  $z = \theta/SE_T$  gives us the exact type I error based on the standard normal distribution, the theoretical type I error of the  $z$ -test based on  $z' = \theta/SE_B$  can be calculated as the following:

$$Type\ I\ error(r) = 1 - \int_{-1.96}^{1.96} \frac{r}{\sqrt{2\pi}} e^{-\frac{(rz')^2}{2}} dz', \quad (14)$$

---

Based on Equation 14, the theoretical type I error of  $z$ -test based on the underestimated standard error from BILOG-MG can be calculated. For example,  $r = 0.5$  indicates that the standard error estimate from BILOG-MG is half of the minimal asymptotic standard error. Then, the  $z$ -statistic is doubled based on Equation 13, and the theoretical type I error becomes 0.32 based on Equation 14.

## 5.2 Results for the Rasch Model

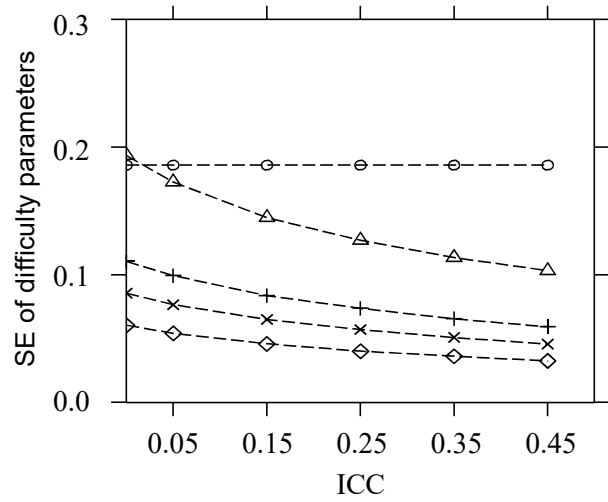
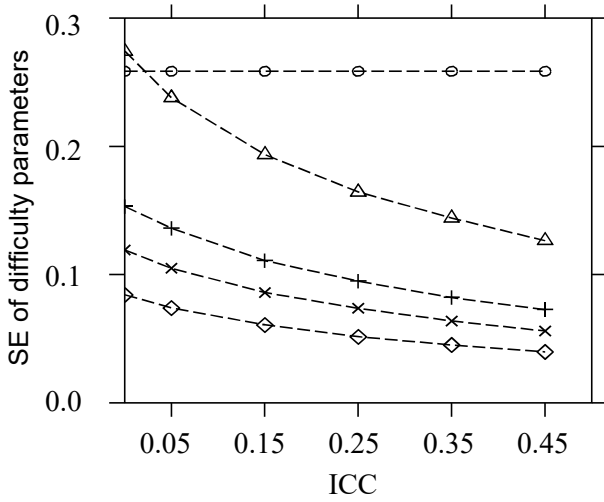
### 5.2.1. Standard Errors of Difficulty Parameters

The standard error estimates of the difficulty parameters in the Rasch model estimated from BILOG-MG are plotted in Figures 1 through 3 to demonstrate the influence of multilevel data structure on the estimation of standard errors. Figure 1, Figure 2, and Figure 3 show the standard error estimates for the cases where the number of groups ( $n_G$ ) are 50, 100, and 200, respectively. Each subplot in the figures shows the standard error estimates for a specific value of item difficulty parameters ( $b$ ), and each line in the subplots shows the standard error estimates for a specific value of group sizes ( $n_W$ ). Because of space limitations, the ratios defined by Equation 11 are presented only for the number of groups ( $n_G$ ) of 50 in Table 1. In the table, the numbers in the parentheses are the type I errors for the corresponding values of  $r$  that were calculated based on Equation 14.

**Figure 1.** Standard error estimates of difficulty parameters in the Rasch model (BILOG, nG=50)

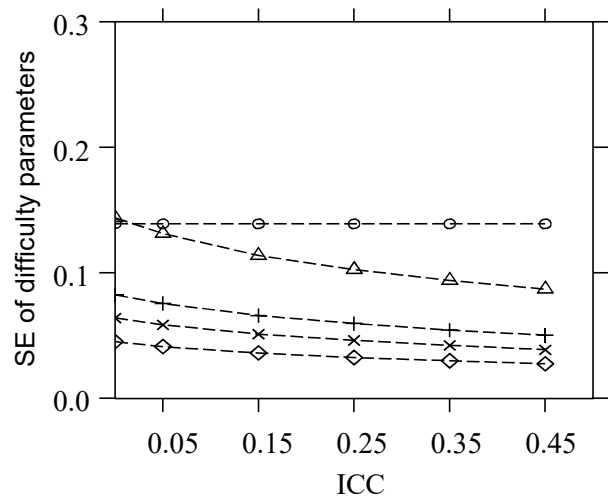
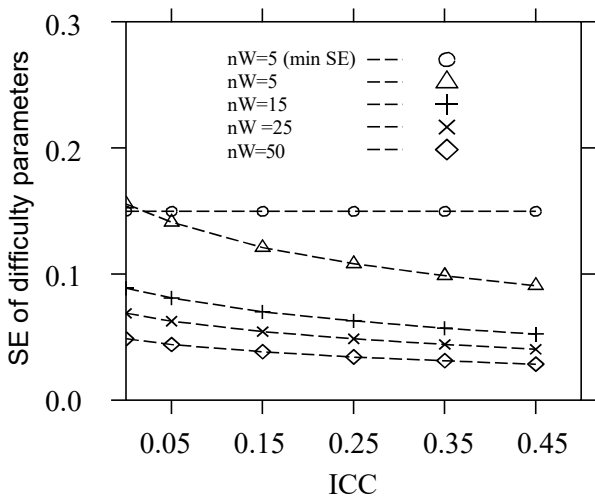
a) Item 1 (b=-3), Item 7 (b=3)

b) Item 2 (b=-2), Item 6 (b=2)



c) Item 3 (b=-1), Item 5 (b=1)

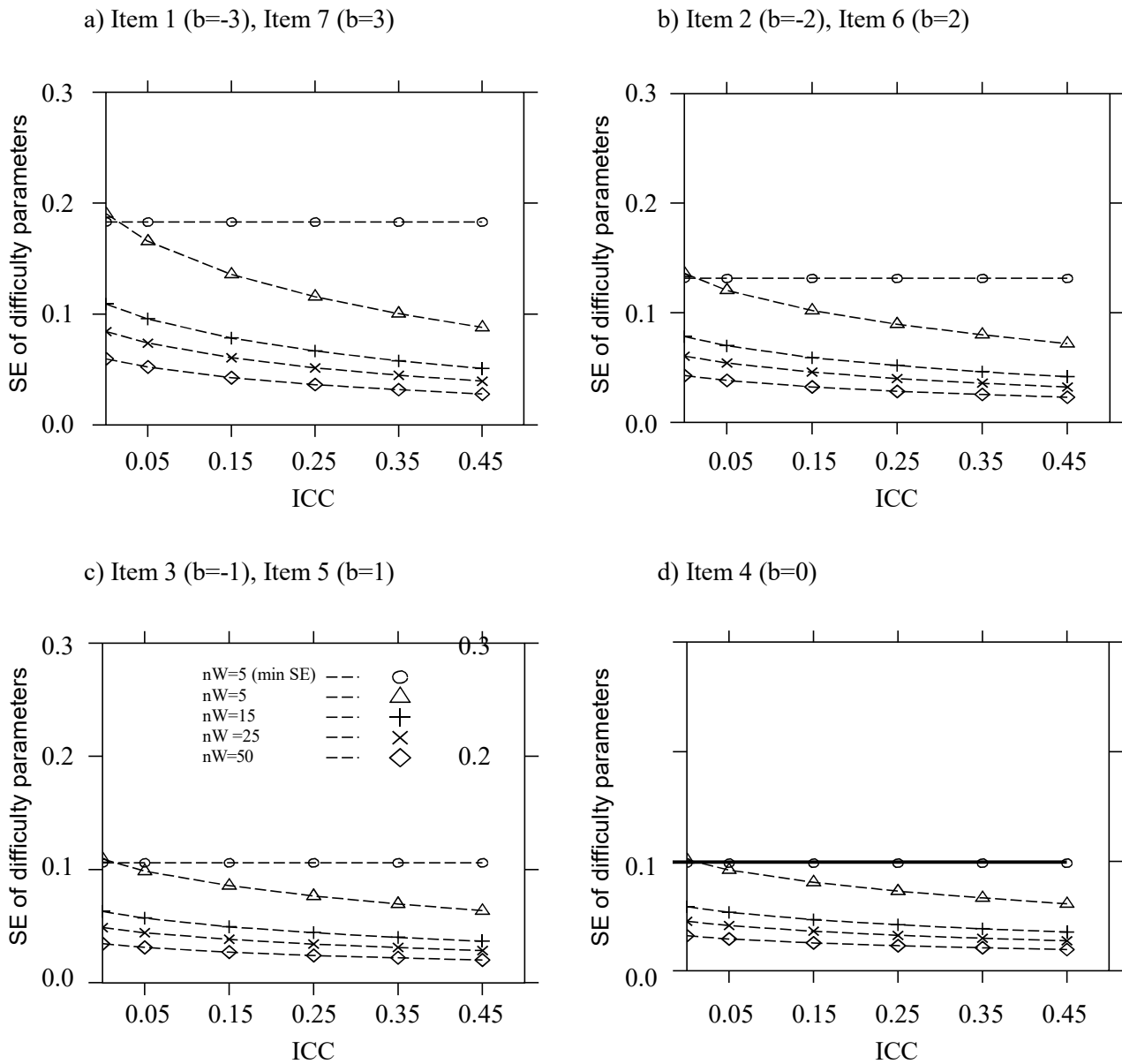
d) Item 4 (b=0)



**Note.** This figure provides a graphical illustration of changes in standard error estimates depending ICC when the number of groups (nG) is 50. For the group size (nW) of 5, the minimal asymptotic standard errors were plotted together for comparison.



**Figure 2.** Standard error estimates of difficulty parameters in the Rasch model (BILOG,  $nG=100$ )

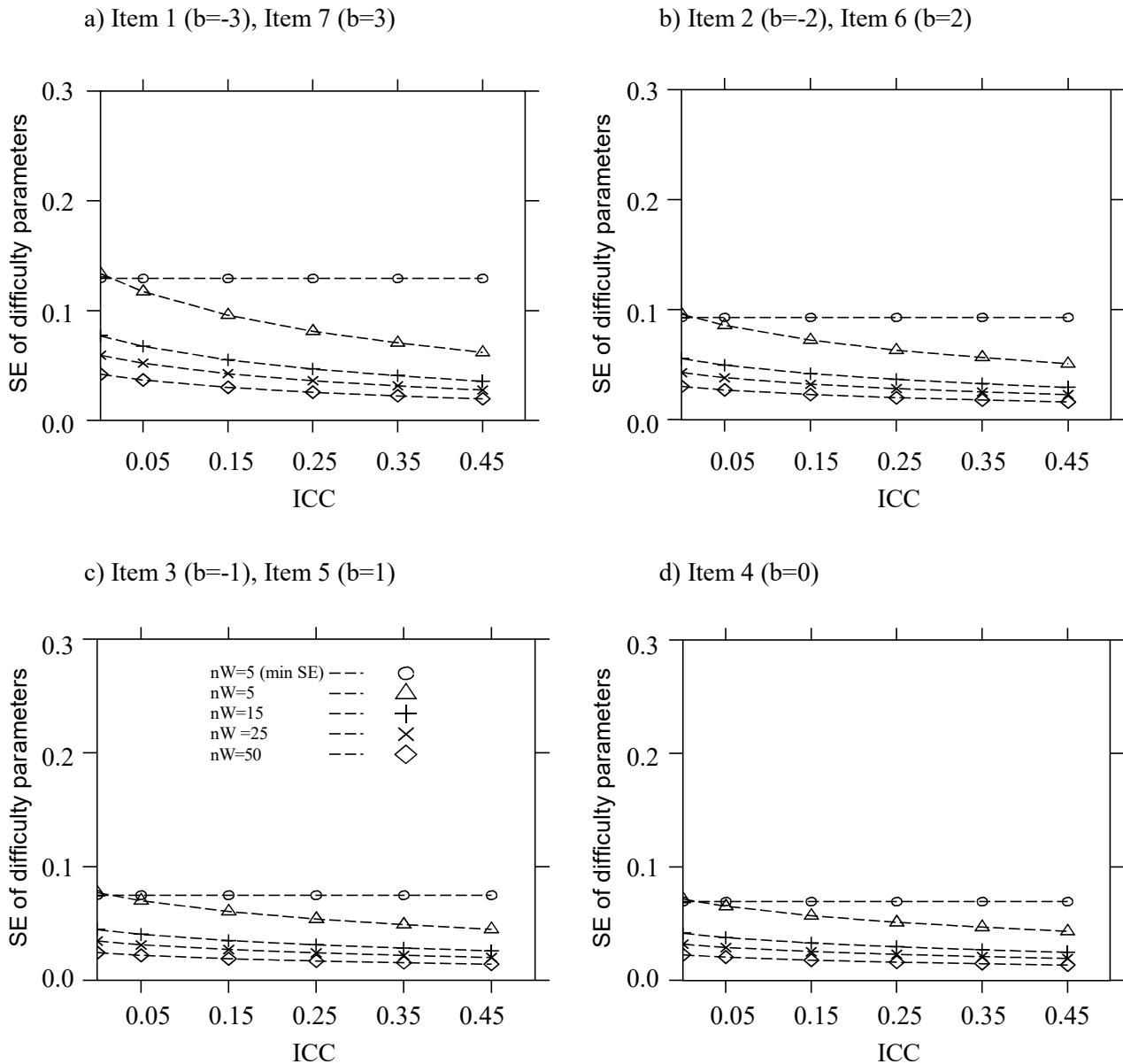


**Note.** This figure provides a graphical illustration of changes in standard error estimates depending ICC when the number of groups ( $nG$ ) is 100. For the group size ( $nW$ ) of 5, the minimal asymptotic standard errors were plotted together for comparison.

Several trends can be identified from the results. Most importantly, the results show that the standard errors estimates decrease as the values of the ICC increase. The decrease is most prominent when the number of groups and the group sizes are small. For example, in [Figure 1a](#), the standard error estimates for  $nG = 50$ ,  $nW = 5$ , and  $b=-3$  or  $3$  decrease from 0.2742 to 0.1265 as the values of the ICC increase from 0 to 0.45. Also, as can be seen from [Table 1](#), the ratio  $r$  for  $ICC = 0.45$ ,  $nG = 50$ ,  $nW = 5$ , and  $b=-3$  or  $3$  was 0.49, which indicates that the standard error estimate from the standard single-level Rasch model, which is 0.1265 in this case, is about half of the minimal asymptotic standard error. Note that the minimal asymptotic standard error for  $nG = 50$ ,  $nW = 5$ , and  $b=-3$  or  $3$  is 0.2586.

Secondly, the effect of the ICC on the estimation for the standard error decrease as the number of groups ( $nG$ ) and the group sizes ( $nW$ ) increase. For example, in Figure 1a, the standard error estimate for  $nW = 5$  decrease more than  $nW = 50$  as the values of the ICC increase. Also, the decrease is more prominent for  $nG = 50$  (Figure 1) than  $nG = 200$  (Figure 3).

**Figure 3.** Standard error estimates of difficulty parameters in the Rasch model (BILOG,  $nG=200$ )



**Note.** This figure provides a graphical illustration of changes in standard error estimates depending ICC when the number of groups ( $nG$ ) IS 200. For the group size ( $nW$ ) of 5, the minimal asymptotic standard errors were plotted together for comparison.

**Table 1.** The Ratios and Type I Errors for the Rasch Model When  $nG = 50$ 

| ICC  | Groups | Groups<br>Sizes | Item 1         | Item 2         | Item 3         | Item 4         | Item 5         | Item 6          | Item 7         |
|------|--------|-----------------|----------------|----------------|----------------|----------------|----------------|-----------------|----------------|
| 0.05 | 50     | 5               | 0.92<br>(0.07) | 0.93<br>(0.07) | 0.94<br>(0.06) | 0.94<br>(0.06) | 0.94<br>(0.07) | 0.93<br>(0.07)  | 0.92<br>(0.07) |
| 0.05 | 50     | 15              | 0.91<br>(0.07) | 0.92<br>(0.07) | 0.94<br>(0.07) | 0.94<br>(0.06) | 0.94<br>(0.07) | 0.92<br>(0.07)  | 0.91<br>(0.07) |
| 0.05 | 50     | 25              | 0.91<br>(0.08) | 0.92<br>(0.07) | 0.94<br>(0.07) | 0.94<br>(0.07) | 0.94<br>(0.07) | 0.92<br>(0.07)  | 0.91<br>(0.08) |
| 0.05 | 50     | 50              | 0.90<br>(0.08) | 0.92<br>(0.07) | 0.93<br>(0.07) | 0.94<br>(0.07) | 0.93<br>(0.07) | 0.92<br>(0.07)  | 0.90<br>(0.08) |
| 0.15 | 50     | 5               | 0.75<br>(0.14) | 0.78<br>(0.13) | 0.81<br>(0.11) | 0.82<br>(0.11) | 0.81<br>(0.11) | 0.78<br>(0.13)  | 0.75<br>(0.14) |
| 0.15 | 50     | 15              | 0.74<br>(0.15) | 0.78<br>(0.13) | 0.81<br>(0.11) | 0.82<br>(0.11) | 0.81<br>(0.11) | 0.78<br>(0.13)  | 0.74<br>(0.15) |
| 0.15 | 50     | 25              | 0.74<br>(0.14) | 0.78<br>(0.13) | 0.81<br>(0.11) | 0.82<br>(0.11) | 0.81<br>(0.11) | 0.78<br>(0.13)  | 0.74<br>(0.14) |
| 0.15 | 50     | 50              | 0.74<br>(0.15) | 0.78<br>(0.13) | 0.81<br>(0.11) | 0.82<br>(0.11) | 0.81<br>(0.11) | 0.784<br>(0.13) | 0.74<br>(0.14) |
| 0.25 | 50     | 5               | 0.64<br>(0.21) | 0.68<br>(0.18) | 0.72<br>(0.16) | 0.74<br>(0.15) | 0.72<br>(0.16) | 0.69<br>(0.18)  | 0.64<br>(0.21) |
| 0.25 | 50     | 15              | 0.64<br>(0.21) | 0.69<br>(0.18) | 0.73<br>(0.15) | 0.74<br>(0.15) | 0.73<br>(0.15) | 0.69<br>(0.18)  | 0.64<br>(0.21) |
| 0.25 | 50     | 25              | 0.64<br>(0.21) | 0.69<br>(0.18) | 0.73<br>(0.15) | 0.74<br>(0.14) | 0.73<br>(0.15) | 0.69<br>(0.18)  | 0.64<br>(0.21) |
| 0.25 | 50     | 50              | 0.63<br>(0.22) | 0.68<br>(0.18) | 0.72<br>(0.16) | 0.74<br>(0.15) | 0.72<br>(0.16) | 0.68<br>(0.18)  | 0.63<br>(0.22) |
| 0.35 | 50     | 5               | 0.56<br>(0.27) | 0.61<br>(0.23) | 0.66<br>(0.20) | 0.68<br>(0.19) | 0.66<br>(0.20) | 0.61<br>(0.23)  | 0.55<br>(0.28) |
| 0.35 | 50     | 15              | 0.55<br>(0.28) | 0.61<br>(0.23) | 0.66<br>(0.20) | 0.68<br>(0.18) | 0.66<br>(0.20) | 0.61<br>(0.23)  | 0.55<br>(0.28) |
| 0.35 | 50     | 25              | 0.55<br>(0.28) | 0.61<br>(0.23) | 0.66<br>(0.19) | 0.68<br>(0.18) | 0.66<br>(0.19) | 0.61<br>(0.23)  | 0.55<br>(0.28) |
| 0.35 | 50     | 50              | 0.55<br>(0.28) | 0.61<br>(0.23) | 0.66<br>(0.19) | 0.68<br>(0.18) | 0.66<br>(0.19) | 0.61<br>(0.23)  | 0.56<br>(0.28) |
| 0.45 | 50     | 5               | 0.49<br>(0.34) | 0.56<br>(0.28) | 0.61<br>(0.23) | 0.63<br>(0.22) | 0.61<br>(0.23) | 0.55<br>(0.28)  | 0.49<br>(0.34) |
| 0.45 | 50     | 15              | 0.49<br>(0.34) | 0.55<br>(0.28) | 0.61<br>(0.24) | 0.63<br>(0.22) | 0.61<br>(0.23) | 0.55<br>(0.28)  | 0.49<br>(0.34) |
| 0.45 | 50     | 25              | 0.49<br>(0.34) | 0.55<br>(0.28) | 0.60<br>(0.24) | 0.62<br>(0.22) | 0.60<br>(0.24) | 0.55<br>(0.28)  | 0.48<br>(0.34) |
| 0.45 | 50     | 50              | 0.48<br>(0.34) | 0.55<br>(0.28) | 0.61<br>(0.24) | 0.63<br>(0.22) | 0.61<br>(0.23) | 0.55<br>(0.28)  | 0.49<br>(0.34) |

*Notes.* For each simulation condition, the numbers in the first line represent ratios  $r$  based on Equation 11, and the numbers in parentheses in the second line represent type I errors based on Equation 14.

### 5.2.2. Biases for Difficulty Parameters

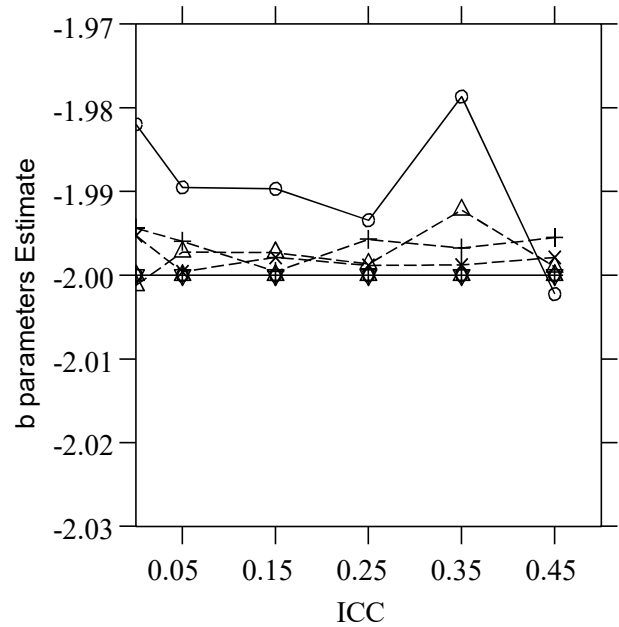
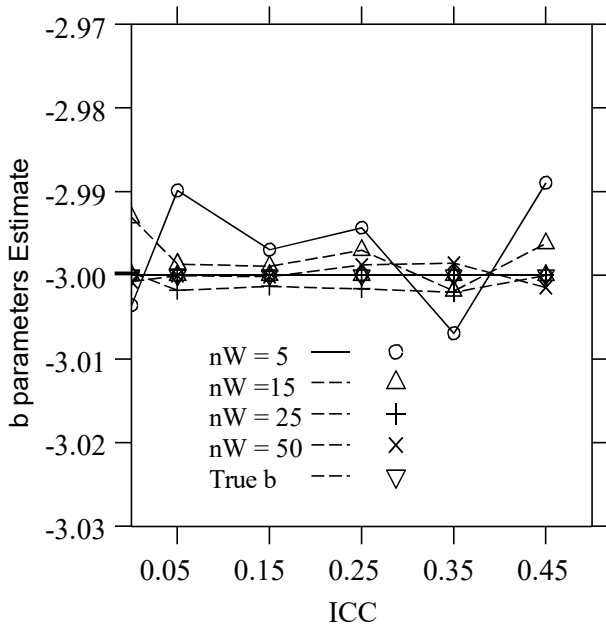
The estimates for the item difficulty parameters were also monitored to check the influence of manipulated factors on the estimation of item difficulty parameters. Because of the space limitations, the parameter estimates are presented only for the number of groups ( $nG$ ) 50 in Figure 4. In contrast to the results of standard error estimates, it seemed that the ICC did not affect the estimation of the item difficulty parameters. The figure does not show any systematic

pattern, and the parameter estimates remain stable across the values of the ICC. Similarly, no systematic pattern was observed for the number of groups ( $nG$ ) 100 and 200.

**Figure 4.** Biases of difficulty parameters in the Rasch model (BILOG,  $nG=50$ )

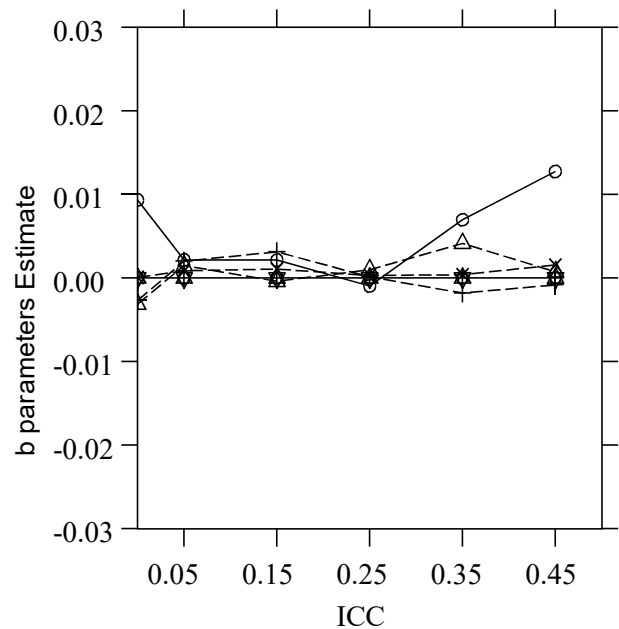
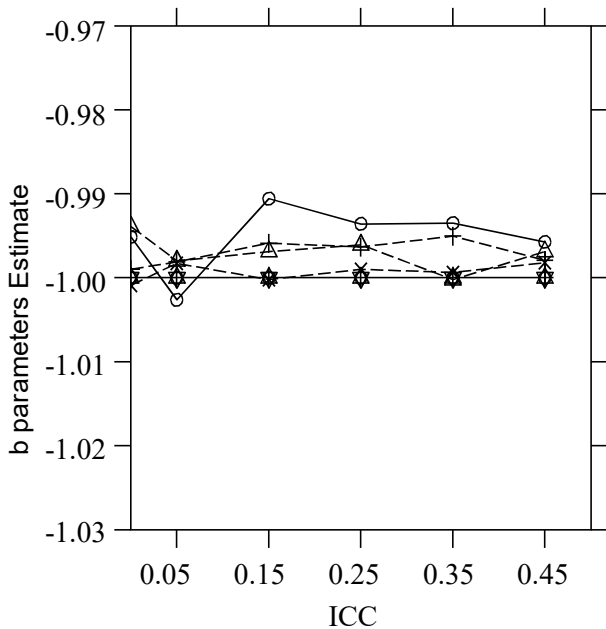
a) Item 1 ( $b=-3$ )

b) Item 2 ( $b=-2$ )



c) Item 3 ( $b=-1$ )

d) Item 4 ( $b=0$ )



**Note.** This figure provides a graphical illustration of changes in bias of parameter estimates depending on ICC for the number of groups ( $nG$ ) 50. The true values of parameters were plotted as horizontal lines.

### 5.3. Results for the 2PL and 3PL

Overall, similar patterns were observed for the 2PL and 3PL models. The standard error estimates decreased as the values of the ICC increased. Also, the estimates for item parameters were stable across different values of the ICC, and no systematic pattern was observed for the bias of parameter estimates. Because of space limitations, only parts of the results are presented in Table 2. The results for the number of groups 100 and 200 also showed similar patterns but are not presented because of the limitation of space.

**Table 2.** The Ratios and Type I Errors for the 2PL and 3PL Models When  $nG = 50$

| ICC  | Groups | Groups Sizes | 2PL (a=1) | 2PL (b=3) | 3PL (a=1) | 3PL (b=3) | 3PL (c=0.2) |
|------|--------|--------------|-----------|-----------|-----------|-----------|-------------|
| 0.05 | 50     | 5            | 0.61      | 0.48      | 0.54      | 0.75      | 0.57        |
|      |        |              | (0.23)    | (0.34)    | (0.28)    | (0.13)    | (0.25)      |
| 0.05 | 50     | 15           | 0.60      | 0.47      | 0.65      | 0.71      | 0.72        |
|      |        |              | (0.23)    | (0.34)    | (0.20)    | (0.16)    | (0.15)      |
| 0.05 | 50     | 25           | 0.60      | 0.46      | 0.69      | 0.72      | 0.79        |
|      |        |              | (0.23)    | (0.35)    | (0.17)    | (0.15)    | (0.11)      |
| 0.05 | 50     | 50           | 0.60      | 0.47      | 0.75      | 0.75      | 0.85        |
|      |        |              | (0.23)    | (0.35)    | (0.13)    | (0.14)    | (0.09)      |
| 0.15 | 50     | 5            | 0.58      | 0.44      | 0.43      | 0.57      | 0.59        |
|      |        |              | (0.25)    | (0.38)    | (0.39)    | (0.26)    | (0.24)      |
| 0.15 | 50     | 15           | 0.60      | 0.42      | 0.50      | 0.60      | 0.74        |
|      |        |              | (0.23)    | (0.40)    | (0.32)    | (0.23)    | (0.14)      |
| 0.15 | 50     | 25           | 0.56      | 0.43      | 0.54      | 0.57      | 0.78        |
|      |        |              | (0.26)    | (0.39)    | (0.28)    | (0.25)    | (0.12)      |
| 0.15 | 50     | 50           | 0.57      | 0.44      | 0.60      | 0.59      | 0.81        |
|      |        |              | (0.25)    | (0.38)    | (0.23)    | (0.24)    | (0.10)      |
| 0.25 | 50     | 5            | 0.53      | 0.40      | 0.34      | 0.53      | 0.60        |
|      |        |              | (0.29)    | (0.43)    | (0.49)    | (0.29)    | (0.23)      |
| 0.25 | 50     | 15           | 0.54      | 0.39      | 0.42      | 0.51      | 0.73        |
|      |        |              | (0.28)    | (0.44)    | (0.40)    | (0.31)    | (0.15)      |
| 0.25 | 50     | 25           | 0.54      | 0.39      | 0.46      | 0.50      | 0.76        |
|      |        |              | (0.28)    | (0.44)    | (0.36)    | (0.32)    | (0.13)      |
| 0.25 | 50     | 50           | 0.54      | 0.38      | 0.50      | 0.48      | 0.77        |
|      |        |              | (0.28)    | (0.44)    | (0.32)    | (0.34)    | (0.13)      |
| 0.35 | 50     | 5            | 0.51      | 0.37      | 0.29      | 0.46      | 0.61        |
|      |        |              | (0.31)    | (0.45)    | (0.56)    | (0.36)    | (0.22)      |
| 0.35 | 50     | 15           | 0.52      | 0.37      | 0.37      | 0.44      | 0.71        |
|      |        |              | (0.30)    | (0.46)    | (0.46)    | (0.38)    | (0.16)      |
| 0.35 | 50     | 25           | 0.52      | 0.36      | 0.40      | 0.42      | 0.72        |
|      |        |              | (0.30)    | (0.47)    | (0.42)    | (0.40)    | (0.15)      |
| 0.35 | 50     | 50           | 0.52      | 0.36      | 0.43      | 0.41      | 0.73        |
|      |        |              | (0.30)    | (0.47)    | (0.39)    | (0.41)    | (0.15)      |
| 0.45 | 50     | 5            | 0.50      | 0.37      | 0.26      | 0.42      | 0.61        |
|      |        |              | (0.32)    | (0.46)    | (0.60)    | (0.40)    | (0.23)      |
| 0.45 | 50     | 15           | 0.52      | 0.36      | 0.33      | 0.40      | 0.69        |
|      |        |              | (0.30)    | (0.47)    | (0.50)    | (0.42)    | (0.17)      |
| 0.45 | 50     | 25           | 0.52      | 0.36      | 0.36      | 0.38      | 0.70        |
|      |        |              | (0.30)    | (0.47)    | (0.47)    | (0.45)    | (0.16)      |
| 0.45 | 50     | 50           | 0.52      | 0.36      | 0.38      | 0.36      | 0.68        |
|      |        |              | (0.30)    | (0.47)    | (0.44)    | (0.47)    | (0.17)      |

*Note.* For each simulation condition, the numbers in the first line represent ratios  $r$  based on Equation 11, and the numbers in parentheses in the second line represent type I errors based on Equation 14.

## 6. SIMULATION STUDY2

### 6.1. Simulation Designs

This simulation study was to design to compare type I error rates of DIF tests using models with and without reflecting multilevel data structure. To do so, data sets were generated using the same setting of the Rasch model in the simulation study 1 with no DIF for a studied item across hypothetical binary groups. In generating data sets, item difficulty parameters were set at (-3, -2, -1, 0, 1, 2, 3), and multilevel structure was implemented with different values of the ICC, which are 0, .05, .15, .25, .35, and .45. Two different kinds of DIF tests were performed across those hypothetical groups using the Lord’s chi-square test and the Rasch model formulated in hierarchical generalized linear model (HGLM). The Lord’s chi-square tests were performed using parameter estimates and their standard errors estimated from BILOG-MG. Also, another DIF tests were performed based on the Rasch model that was formulated in the hierarchical generalized linear model (HGLM) in which items, persons, and groups are modeled in Level-1, Level-2, and Level-3 respectively.

### 6.2. Results

The results of DIF tests using Lord’s chi-square tests and the multilevel Rasch model for the number of groups ( $nG$ ) 50 are presented in Table 3. In the table, the numbers in the first line of each simulation condition represent type I errors from the Lord’s chi-square tests, and the numbers in parentheses in the second line represent type I errors from the multilevel Rasch model. From the table, it can be seen that the type I error rates of the Lord’s chi-square tests are inflated up to .270 as the values of the ICC increase, whereas the type I error rates of the multilevel Rasch model remain quite stable close to the nominal level of significance, which is .05.

**Table 3.** DIF using Lord Chi square Test vs HGLM When  $nG = 50$

| ICC  | Groups | Groups Sizes | Item 1            | Item 2           | Item 3           | Item 4           | Item 5           | Item 6           | Item 7           |
|------|--------|--------------|-------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| 0.05 | 50     | 5            | 0.005<br>(0.026)  | 0.044<br>(0.025) | 0.082<br>(0.051) | 0.064<br>(0.054) | 0.098<br>(0.053) | 0.048<br>(0.039) | 0.008<br>(0.022) |
| 0.05 | 50     | 15           | 0.009<br>(0.027)  | 0.033<br>(0.028) | 0.075<br>(0.050) | 0.078<br>(0.046) | 0.066<br>(0.034) | 0.066<br>(0.026) | 0.005<br>(0.021) |
| 0.05 | 50     | 25           | 0.002<br>(0.0274) | 0.043<br>(0.031) | 0.056<br>(0.053) | 0.042<br>(0.053) | 0.058<br>(0.048) | 0.035<br>(0.033) | 0.002<br>(0.022) |
| 0.05 | 50     | 50           | 0.003<br>(0.023)  | 0.046<br>(0.035) | 0.062<br>(0.053) | 0.073<br>(0.064) | 0.068<br>(0.074) | 0.049<br>(0.055) | 0.007<br>(0.039) |
| 0.15 | 50     | 5            | 0.021<br>(0.052)  | 0.116<br>(0.081) | 0.136<br>(0.093) | 0.130<br>(0.080) | 0.114<br>(0.060) | 0.105<br>(0.053) | 0.026<br>(0.027) |
| 0.15 | 50     | 15           | 0.012<br>(0.034)  | 0.072<br>(0.039) | 0.100<br>(0.076) | 0.112<br>(0.075) | 0.093<br>(0.054) | 0.080<br>(0.049) | 0.013<br>(0.042) |
| 0.15 | 50     | 25           | 0.022<br>(0.051)  | 0.081<br>(0.061) | 0.121<br>(0.066) | 0.143<br>(0.073) | 0.129<br>(0.051) | 0.096<br>(0.041) | 0.031<br>(0.039) |
| 0.15 | 50     | 50           | 0.021<br>(0.035)  | 0.094<br>(0.035) | 0.131<br>(0.059) | 0.134<br>(0.055) | 0.121<br>(0.075) | 0.072<br>(0.063) | 0.015<br>(0.045) |
| 0.25 | 50     | 5            | 0.028<br>(0.022)  | 0.094<br>(0.025) | 0.112<br>(0.041) | 0.130<br>(0.055) | 0.129<br>(0.042) | 0.105<br>(0.045) | 0.020<br>(0.030) |
| 0.25 | 50     | 15           | 0.032<br>(0.037)  | 0.121<br>(0.044) | 0.167<br>(0.044) | 0.149<br>(0.040) | 0.130<br>(0.034) | 0.109<br>(0.044) | 0.048<br>(0.046) |
| 0.25 | 50     | 25           | 0.028<br>(0.030)  | 0.132<br>(0.040) | 0.159<br>(0.047) | 0.164<br>(0.041) | 0.145<br>(0.029) | 0.127<br>(0.041) | 0.047<br>(0.029) |



**Table 3.** Continues

|      |    |    |                  |                  |                  |                  |                  |                  |                  |
|------|----|----|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| 0.25 | 50 | 50 | 0.038<br>(0.035) | 0.143<br>(0.046) | 0.163<br>(0.055) | 0.164<br>(0.045) | 0.162<br>(0.041) | 0.122<br>(0.026) | 0.045<br>(0.034) |
| 0.35 | 50 | 5  | 0.069<br>(0.035) | 0.141<br>(0.045) | 0.193<br>(0.031) | 0.200<br>(0.043) | 0.213<br>(0.048) | 0.155<br>(0.035) | 0.068<br>(0.027) |
| 0.35 | 50 | 15 | 0.049<br>(0.027) | 0.138<br>(0.040) | 0.176<br>(0.047) | 0.176<br>(0.052) | 0.186<br>(0.054) | 0.133<br>(0.042) | 0.055<br>(0.033) |
| 0.35 | 50 | 25 | 0.065<br>(0.036) | 0.178<br>(0.041) | 0.195<br>(0.043) | 0.228<br>(0.034) | 0.227<br>(0.044) | 0.166<br>(0.029) | 0.088<br>(0.030) |
| 0.35 | 50 | 50 | 0.090<br>(0.043) | 0.163<br>(0.044) | 0.227<br>(0.050) | 0.230<br>(0.037) | 0.213<br>(0.033) | 0.167<br>(0.038) | 0.063<br>(0.031) |
| 0.45 | 50 | 5  | 0.182<br>(0.047) | 0.316<br>(0.078) | 0.353<br>(0.083) | 0.341<br>(0.063) | 0.317<br>(0.047) | 0.317<br>(0.045) | 0.197<br>(0.031) |
| 0.45 | 50 | 15 | 0.118<br>(0.052) | 0.253<br>(0.051) | 0.281<br>(0.049) | 0.258<br>(0.043) | 0.258<br>(0.051) | 0.236<br>(0.067) | 0.120<br>(0.038) |
| 0.45 | 50 | 25 | 0.137<br>(0.032) | 0.231<br>(0.044) | 0.268<br>(0.041) | 0.298<br>(0.038) | 0.282<br>(0.043) | 0.238<br>(0.034) | 0.116<br>(0.034) |
| 0.45 | 50 | 50 | 0.131<br>(0.039) | 0.204<br>(0.048) | 0.259<br>(0.085) | 0.270<br>(0.070) | 0.263<br>(0.049) | 0.213<br>(0.026) | 0.137<br>(0.034) |

*Note.* For each simulation condition, the numbers in the first line represent type I errors from Lord Chi Square tests, and the numbers in parentheses in the second line represent type I errors from HGLM.

## 7. DISCUSSION

It is well known that applying single-level statistical models to multilevel data may produce underestimated standard error estimates, which in turn result in invalid statistical inferences based on such underestimated standard errors. The goal of this study was to examine the impact of multilevel data structure on the estimation of standard errors in dichotomous IRT models in order to explicitly demonstrate the degree of such underestimation in IRT. Given existing and potential IRT applications in which standard error estimates for item parameters play a crucial role (Toland, 2008), it is important to understand the behavior of the standard error estimation of the IRT models in the presence of multilevel data. Our simulation study showed that the degree of underestimation could be quite huge depending on the values of the ICC. In the most severe case, where the value of the ICC was .45, the standard error estimate from

BILOG-MG was about half of the minimal asymptotic standard error; the type I error rates of the Lord's chi-square tests were inflated up to .35; and the type I error rates of hypothetical  $z$ -test using Equation 14 were also inflated up to .47. However, the type I error rates of DIF tests using the multilevel Rasch model were close to the nominal level of  $\alpha$ , which is .05. Multilevel data structure did not affect item parameter estimates.

The results of this study match those of previous studies. Ignoring multilevel data structure caused underestimated standard errors in regression (Goldstein, 1987) and SEM (Satorra & Muthen, 1995). Barcikowski (1981) also found that even a small amount of the ICC can produce dramatic increases in the actual type I error of a  $t$ -test. For example, with the group size of 50, an ICC of .05, which is usually considered small, produced a type I error of .30. In IRT, Finch and French (2011) showed that the type I error of a DIF test using a standard logistic regression can be inflated in the presence of multilevel data structure. In their work, the type I error rate was inflated up to .44 when the value of the ICC was .45. Because the reason for such type I error inflation is the underestimated standard errors, in this study, we wanted to explicitly show the degree of underestimation in IRT settings.

The underestimation of standard errors is caused by the violation of the independent assumption of traditional statistical models. In the presence of multilevel data structure, individuals share

common experiences due to closeness in space or time, which makes individuals within the same context more similar to one another. Therefore, observed variance within clusters is typically less than observed variance between clusters. When observations are assumed to be independent, overall variance is calculated without considering the similarity among individuals within clusters, and tends to be underestimated. In fact, as the values of ICC increase, the standard error estimates should increase if the multilevel data structure is properly handled by statistical models (Snijders & Bosker, 1999; Raudenbush, 1997).

Taken all together, the results of this study suggest that ignoring multilevel data structure in the estimation of IRT models could result in underestimated standard errors for item parameter estimates. More importantly, the extents to which standard errors are underestimated are quite huge. Many evidences from previous studies also suggest that standard error estimates in statistical models in general are quite sensitive to multilevel data structure. Therefore, ignoring multilevel data structure could result in invalid statistical inferences in IRT settings. Therefore, researchers who want to use IRT applications in which standard error estimates of item parameters play a crucial role need to check whether their data sets have multilevel data structure or not. In the presence of multilevel structure, traditional single level model could be problematic. Instead, multilevel IRT models are recommended.

## ORCID

Hyung Rock Lee  <https://orcid.org/0000-0002-7415-9466>

Sunbok Lee  <https://orcid.org/0000-0020-0924-7056>

Jaeyun Sung  <https://orcid.org/0000-0001-7461-3123>

## 8. REFERENCES

- Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational and Behavioral Statistics*, 6, 267–285.
- De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guildford Press.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Mahwah, NJ: Erlbaum.
- Finch, W. H., & French, B. F. (2011). Estimation of mimic model parameters with multilevel data. *Structural Equation Modeling*, 1, 229–252.
- Goldstein, H. (1987). *Multilevel statistical models*. London: Edward Arnold.
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60–87.
- Hox, J. (1998). Multilevel modeling: When and why. In *Classification, data analysis, and data highways* (pp. 147–154). Springer.
- Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, 49, 82–100.
- Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling*, 8, 325–352.
- Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38, 79–93.
- Kamata, A., & Vaughn, B. K. (2011). Multilevel IRT modeling. *Handbook of advanced multilevel analysis* (pp. 41-57). New York, NY: Taylor and Francis Group.
- Kim, S.-H., & Cohen, A. S. (1995). A comparison of lord's chi-square, raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, 8, 291–312.
- Kish, L. (1965). *Survey sampling*. New York: Wiley.

- Lord, F. M. (1980). *Applications of item response to theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1, 86–92.
- Oshima, T., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (dfit) framework. *Journal of Educational Measurement*, 43, 1–17.
- R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org/>
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2, 173.
- Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer.
- Satorra, A., & Muthen, B. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25, 267–316.
- Snijders, T. A., & Bosker, R. J. (1999). *Introduction to multilevel analysis*. London: Sage.
- Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publishers.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.
- Tate, R. L., & Wongbundhit, Y. (1983). Random versus nonrandom coefficient models for multilevel analysis. *Journal of Educational and Behavioral Statistics*, 8, 103–120.
- Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, 47(4), 397–412.
- Toland, M. D. (2008). *Determining the accuracy of item parameter standard error of estimates in bilog-mg 3*. ProQuest.
- Veerkamp, W. J., & Glas, C. A. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, 25, 373–389.
- Wright, B., & Stone, M. (1979). *Best test design: A handbook for rasch Measurement*. Chicago: MESA.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). Bilog-mg: Multiple-group IRT analysis and test maintenance for binary items. *Chicago: Scientific Software International*, 4, 10.

## Development of Exposure to English Scale and Investigation of Exposure Effect to Achievement

Mustafa Gökcan <sup>1,\*</sup> Derya Çobanoğlu Aktan <sup>1</sup>

<sup>1</sup> Department of Educational Measurement and Evaluation, Hacettepe University, Ankara, Turkey

### ARTICLE HISTORY

Received: 16 September 2018

Revised: 7 February 2019

Accepted: 16 February 2019

### KEYWORDS

Exposure to English,  
Language Acquisition,  
Structural Equation Modeling,  
Scale Development,  
English Achievement

**Abstract:** An absence of a scale for measuring exposure to the English language, which has a significant effect on English achievement, was detected in the literature. For this reason, in this study, a six-dimensional scale was developed to detect the level of English language exposure and its construct validity was tested. The factor structure of the scale was determined by exploratory factor analysis with the data collected from 784 university students, 726 of whom are undergraduate and 58 of whom are Master's and Ph.D. students. Confirmation of the factor structure of the scale was carried out with a measurement model specified in a structural equation model. A structural equation modeling study was performed along with 233 students from English preparation classes at a university. In the structural model, the effect of exposure to English on the students' scores received from writing in English, speaking in English and the total score (grammar, vocabulary, reading and listening scores) was examined. It was found that exposure to English has a significant effect on all of the three variables. Exposure to English explained the variance of the speaking variable most, while that effect is the least for the writing variable.

## 1. INTRODUCTION

It is a fact that the significance of speaking a language is indisputable and so it is at the heart of life. We use language for expressing our feelings, for achieving our goals and even just for pleasure. Some people do all of these things not with a single language but with two or more languages. Even we can say that now monolingual people are one of the endangered species in most countries of the world. A second language affects people's careers, their future, their ongoing lives, and even their identities. For this reason, it is an important duty for educators to make the process of second language learning more efficient and easier for language learners (Cook, 2008). For this purpose, in foreign language education, theories have been put forward and various methods based on these theories were applied. During the times when there were significant methodological shifts and when the methods based on memorizing grammar rules were replaced with the ones focused on meaning, Stephen Krashen's ideas and the hypotheses

**CONTACT:** Mustafa Gökcan ✉ [gokcan.m@gmail.com](mailto:gokcan.m@gmail.com) 📍 Department of Educational Measurement and Evaluation, Hacettepe University, Ankara, Turkey

have been very influential and methods grounded on Krashen's suggestions have been developed and employed widely around the world (Lightbown & Spada, 2006). As Ellis (2015) noted, Krashen's language acquisition theory had an influential effect on language pedagogy and frequently referred in the books serving as guides for English language teachers. By this means, most of the English language teachers have had the chance to know him.

The importance of Krashen in teaching English as a second or foreign language can be understood from following notes of the Cook. Cook (2008) speaks about two geographical separations of teaching conversational skills in English language teaching. In teaching methods originated from the UK, from the very beginning of the course, speaking the language is demanded besides listening to it. However, in language education systems based on the US or more precisely based on Krashen, it is given importance to listening without speaking. Krashen (2009) emphasizes that speaking English is not a skill to be learned but an outcome emerging by itself after being exposed to an ample amount of comprehensible input. Krashen, in the area of language teaching, is one of the linguists that does not give too much importance to language production. Despite widely being criticized for this reason, Krashen's ideas achieved a significant breakthrough in the studies on language acquisition and simply became a turning point in the area (Mitchell, Myles, & Marsden, 2013). The issue that is mostly emphasized in Krashen's own language acquisition theory is the comprehensible input. That is to say, for language acquisition, the most important factor is the amount of input that the learner confronted during the learning process. Those who are more exposed to English and have frequent language exposure acquire the language more easily (Gökcan & Çobanoğlu Aktan, 2018).

According to the literature, exposure to a language, without any doubt, is a vital ingredient in the learning of any language. Along these lines, Harmer (2007) states that fact as, "As far as we can see, children are not taught language, nor do they set out to learn it consciously. Rather they acquire it subconsciously as a result of the massive exposure to it" (p. 49).

Exposure to language has that same importance in second language acquisition too. The role of exposure is emphasized under the name of "comprehensible input" in Stephen Krashen's theory of language acquisition. (Krashen, 1982). His theory consists of five main hypotheses namely, the acquisition-learning hypothesis, the natural order hypothesis, the monitor hypothesis, the input hypothesis, and the affective filter hypothesis. In the first hypothesis, he emphasizes the distinction between learning a language and acquiring a language. Acquiring a language is a subconscious process, and it is the process we undergo while acquiring our mother tongue. According to Krashen, that process is also possible in second language acquisition, once the individuals get an ample amount of comprehensible input and they focus on meaning rather than form as they do when they acquire their first language. In the second hypothesis, some research findings, indicating that there is a predictable and natural order throughout the acquisition of the grammatical structures and the rules of English, are given. In the monitor hypothesis, Krashen states the role of language learning in the context of second language acquisition. He claims that the learned part of language has an influence on our acquired knowledge, and the former one edits or monitors the utterances initiated by the latter one. In the fourth hypothesis, the input hypothesis, the focus is on enabling the second language learners to get sufficient amount of comprehensible input that is one step beyond their current level of language proficiency. In his last hypothesis, he defines the roles of some affective variables such as attitude and anxiety in second language acquisition context. A filter called "Affective filter", which can be defined as a mental block formed by the negative attitude towards English and high foreign language anxiety may hinder the acquisition of the comprehensible input. In summary, the focus of Krashen's theory of language acquisition is



exposure to the target language. By exposure to a sufficient amount of comprehensible input and low affective filter, one can acquire the target language successfully.

Coupled with the theoretical background supporting the important role of exposure in second language acquisition, in the literature, there are also a number of studies investigating the effect of language exposure on English achievement. Few of them will shortly be reviewed here.

### **1.1. Review of Literature**

Olsson (2012) found that there is a significant positive correlation between exposure to English and English course grades of Swedish 9<sup>th</sup> graders and their scores from the writing section of a national English exam. Djigunović, Nikolov, and Ottó (2008) compared the English achievement of Croatian and Hungarian students. In their study, as the indicator of success in language learning, they chose an English exam, which has for sub-tests as reading comprehension, listening comprehension, speaking, and writing. They looked at whether there are differences in the mean scores obtained from all the subtests between the two countries. Croatian students were found more successful and it was observed that the factors like starting language learning early, more hours for English course, a classroom with fewer students, which are presented as the keys for success in language acquisition, are in fact not that effective. It was stated that the real reason that made the Croatian students more successful than Hungarian students is actually Croatian students' higher level of exposure to English. Derwing, Munro, and Thomson (2007) worked with two groups of elite Canadian immigrants whose jobs vary from doctors to engineers and scientists. Each group has sixteen members, in one group their mother tongue is Mandarin, and in the other group, it is Slavic languages. Over the course of two years, the data was collected at certain intervals and it was investigated whether there was a change in their listening to English, listening comprehension and speaking accuracy. While increases were observed among the Slavic group which was also found to be exposed to English more, no increase was seen among the Mandarin-speaking group. Wolf, Smit, and Lowie (2017) investigated the effect of starting learning English earlier on oral fluency. They found that although starting earlier has an effect, exposure to English outside the classroom is a more effective factor. In his very recent study, Peters (2018) also found that the effect of exposure to English on English achievement is more than the effect of length of the instruction.

In the above-mentioned studies to determine the level of language exposure, including Derwing et al. (2007), all the researchers used different questionnaires without any reliability and validity studies. Derwing et al. (2007), however, just used a limited questionnaire, which measures exposure outside the classroom. In their study (Gökcan & Çobanoğlu Aktan, 2016) developed a scale to measure the English exposure levels for elementary students. However, there is no scale to measure university students' exposure to the English language. In this study, developing a scale for university students is aimed.

As it is seen in the aforementioned studies, despite the fact that the effect of exposure to English on general English achievement or on its components like speaking, writing and reading or listening comprehension was investigated separately, its effect on writing, speaking, reading and listening comprehension in English has not been examined at once. For this reason, in this study firstly a scale that will measure the level of language exposure of university students is developed. Then the effect of exposure on speaking, writing and the total score (an examination that includes questions related to reading and listening comprehension, grammar and vocabulary knowledge) is analyzed with structural equation modeling. Although the use of SEM which is an advanced statistical method has increased recently in the language acquisition studies (Winke, 2014), it is still very rare when compared to other analysis methods (Hancock & Schoonen, 2015). The fact that there are many complex variables in the process of language acquisition requires robust statistical methods like SEM (Winke, 2014). In this study, by



employing SEM, first of all, the factor structure of the exposure scale was confirmed and then the effect of the exposure on writing, speaking and total scores of the university students was investigated with a structural model.

## 2. METHOD

The purpose of this study is to develop a scale to measure Exposure to English and to investigate the effect of language exposure to language achievement. In the study, first of all, item pool was written based on literature and previous studies, and the exploratory factor analysis (EFA) was conducted for the construct validity of the scale. The confirmation of the factor structure obtained after EFA was carried out with a measurement model specified in a structural equation modeling analysis with data collected from a different sample. Then a structural model was specified to investigate the effect of exposure to English achievements of the university students. The analysis of EFA was done with SPSS 23 and the analysis of SEM was performed with Mplus 7. The details of the analysis are explained after the specifications of the participants.

The data for this study were collected in two separate times. The first data set was used in the scale development and the second one used in the structural modeling part. The participants of the first part of the study were 810 (363 male, 447 female) university students, 750 of whom are undergraduate and 60 of whom are Master's and Ph.D. students. They were aged between 18 and 32 years and enrolled in various faculties of a private university (but mainly in the faculty of law, economics and administrative sciences, and engineering). The students who participated in the second part of the research were 247 students receiving their English preparation in a state university.

In the process of developing the exposure to English scale firstly 27 scale items that represent the possible sources and the ways from which the students are considered to get comprehensible input were written by reviewing the related literature and Gökcan and Çobanoğlu Aktan's (2016) scale for exposure to English. This scale for elementary students has five sub-factors (i.e. exposure through friends, school, text, media, software). The reliability coefficients (Cronbach's alpha) of the factors were reported as .901, .889, .769, .741, .765 respectively and the scale consisted of twenty items. In addition to the items that are found in elementary student exposure scale, items related to exposure through English-speaking foreigners were added in the current study. It was thought that sources of exposure are different for elementary students than university students. Those students are most likely to have experiences like traveling abroad or talking to tourists.

To provide evidence related to content validity, the first form of the scale was presented to five experts reviewers (two English teachers, one expert who has a Ph.D. in foreign language education and two experts who have Ph.D. in the program of measurement and evaluation in education). According to the experts, items of the scale were appropriate, thus this first form given in **Appendix 1** was administered to 810 university students without any modification.

The EFA study was carried out with those data and after the analysis, the second form of the scale given in **Appendix 2** was obtained. Then by using the second form, a study of structural equation modeling in which the construct of exposure to English was handled as an exogenous variable was performed. As endogenous variables, speaking, writing and booklet scores of 247 students receiving their English preparation in a university were included in the model.

### 2.1. Data Analysis

Before conducting exploratory factor analysis, the suitability of data for factor extraction was examined. For this purpose, Kaiser-Meyer-Olkin (KMO) to examine the sample size adequacy for factor analysis and the Bartlett sphericity test (which shows that the data significantly differs

from the identity matrix and the data belongs to multivariate distribution) were calculated. The data with missing values were excluded from the analysis. According to Tabachnick and Fidell (2013), if missing values are distributed randomly, observed in different variables and few in number when compared to the complete data (<5%), excluding the observations including missing data won't cause a problem. The analysis was carried out with 784 students after the answers including missing data were excluded from the study. The total scale scores of the students were converted to z values in order to detect if there are any outliers. According to Tabachnick and Fidell (2013), if the latent constructs that the items on which EFA is performed will form are foreseen beforehand, Principal axis factoring should be preferred and if the factors are expected to have statistically significant correlations between each other oblique rotation methods should be applied. In addition to this, Brown (2015) states that oblique rotation provides a more realistic representation of factors. Even if the factors are not correlated, according to Brown oblique rotation will produce the same results by orthogonal rotation. On the contrary, when the factors are correlated, the oblique rotation will produce more accurate results.

The Kaiser criterion was employed for determining the number of important factors. According to the Kaiser criterion, the factors having eigenvalues higher than 1 are regarded as important ones and those whose values are below 1 are not taken into consideration. Moreover, this method is suggested for determining the factor structures of the scales that have 20-50 variables (Alpar, 2011). In addition to the Kaiser criterion, scree plot of the eigenvalues was examined to determine the number of the factors.

The reason behind our choice of Kaiser criterion is the will to represent the sources of exposure to English separately. The item removal process was carried out based on the recommendations in the literature (Comrey & Lee, 1992; Çokluk, Şekercioglu & Büyüköztürk, 2014; Thompson, 2004). Firstly, the items which did not load significantly on any factor were removed. Then the items with factor loadings less than .50 were deleted one by one. And lastly, the complex items that cross-load too highly (e.g., > .32) on two factors were also removed.

To determine the internal consistency of the Exposure to English Scale, Cronbach alpha coefficient was calculated for each sub-factors.

In the second part of the study, before starting structural equation modeling, the data were screened and the assumptions for multivariate statistics were tested. The univariate outliers and Mahalanobis distance were checked. The SEM analysis was conducted with the data of 233 students. The assumptions of linearity multicollinearity, univariate and multivariate normality were examined.

The structural model in this study could be described as a partially latent structural regression model (Kline, 2016), because every variable in its structural part is not latent with multiple indicators. In our model, exposure to English is a latent construct measured by multiple indicators, but the variables of speaking, writing, and booklet are single-indicator measurements. Four model-fit measures which are recommended by Kline (2016) were used to assess the model's overall goodness-of-fit: the ratio of Chi-square ( $\chi^2$ ) to degrees-of-freedom (d.f.); comparative fit index (CFI); root mean square error of approximation (RMSEA); and standardized root mean square residual (SRMR).

### **3. FINDINGS**

#### **3.1. Findings related to EFA**

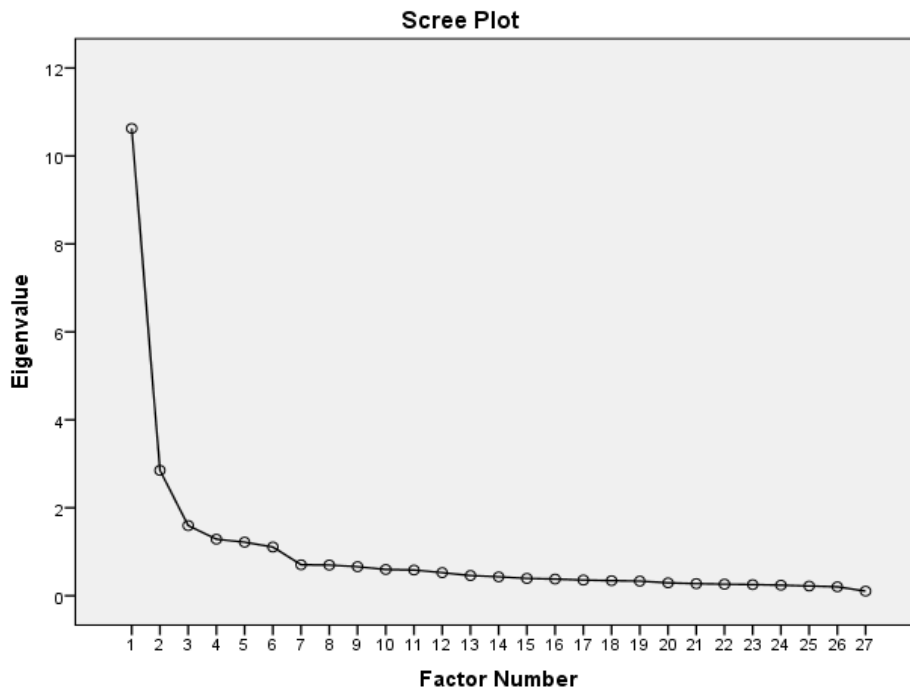
Before conducting exploratory factor analysis, the Kaiser-Meyer-Olkin (KMO) and Bartlett sphericity tests results given in [Table 1](#) were examined. The value of KMO .935 indicated that the sample size is adequate for factor analysis. The Bartlett sphericity p-value for the test was

below .05 and this shows that the data significantly differs from the identity matrix and the data belongs to multivariate distribution. According to these results, the data collected for this study is suitable for factor analysis (Can, 2014; Çokluk et al., 2014;).

The data with missing values were less than %5 of the total data, therefore, they were excluded from the study. The factor structure of the scale was examined with the data of 784 students. The converted z scores of the data indicated that there were no outliers because all the z scores were between -4 and +4.

**Table 1.** KMO and Bartlett's Test

| Name of the test                                 | Value  |
|--|--------|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | .935   |
| Approx. Chi-Square                               | 13.248 |
| Bartlett's Test of Sphericity                    | df     |
|  | .351   |
|  | Sig.   |
|  | .000   |



**Figure 1.** Scree Plot

For the factor extraction, principal axis factoring was performed. Scree plot and Kaiser criterion were used to determine the number of factors. As it is seen from the scree plot in Figure 1, there are 6 points above the point where the curve starts to flatten. Furthermore, there are six factors with eigenvalues higher than the one. This shows that the number of the factors to be extracted should be determined as six (Costello & Osborne, 2005; Thompson, 2004).

Six factors obtained as a result of Kaiser criterion were thought to be fruitful because it provided factors to represent the sources of exposure to English separately. The item removal was done according to the literature (Comrey & Lee, 1992; Çokluk et al. 2014; Thompson, 2004). The items, which did not load significantly on any factor, and the items with factor loadings less than .50 and cross-loaded items were removed. Consequently, five items were deleted and we obtained a scale with 22 items. In Table 2, the eigenvalues of the six factors and the variances explained by the factors are given.

**Table 2.** Initial Eigenvalues and the Total variance explained by the six factors.

| Factor | Eigenvalue | % of Variance | Cumulative % |
|--------|------------|---------------|--------------|
| 1      | 8.251      | 37.505        | 37.505       |
| 2      | 2.592      | 11.781        | 49.286       |
| 3      | 1.485      | 6.751         | 56.038       |
| 4      | 1.276      | 5.800         | 61.837       |
| 5      | 1.102      | 5.009         | 66.847       |
| 6      | 1.053      | 4.786         | 71.632       |

The six factors extracted by EFA were named as Text, School, Media, Friends & Family, Computer, and Foreigners. Although one factor consists of just two items, it was retained in the scale. This is because it is stated that, in multidimensional scales, if the factor loadings of the two items are high and there is no difficulty in interpreting and naming the factor, the factor including two items may not be removed from the scale (Worthington & Whittaker, 2006). The suggestions in the literature for having at least three indicators per factor are found under the title of model identification. The CFA models may be under-identified, just-identified or over-identified, and the parameters of the model can only be estimated when the model is over-identified. In CFA models, the degrees of freedom (df) equal to the number of parameters in the input matrix minus the number of unique free parameters which are estimated from that matrix, and the model is over-identified when the *df* for the model is positive. In other words, to be able to get over-identified CFA models, the number of parameters in the input matrix should be more than the number of freely estimated parameters of the CFA model. While the parameters of the input matrix are the variances of the indicators and the covariances between them, the parameters of the CFA model to be freely estimated are the factor loadings, factor variances, and covariances, error variances and covariances of the indicators etc. If there is only one dimension, the latent construct should be measured by at least three observed variables to meet the conditions of identification. However, if the scale consists of more than one dimension, models, which include two indicators per factor, can also be over-identified. There will be a problem of empirical under identification if the correlations between the factors are equal to 0. However, if the factors are correlated then the model won't have an identification problem, and the parameters of the CFA model can be estimated with ease (Brown, 2015; Tabachnick & Fidell, 2013). In our case, the scale is multi-dimensional and there are significant correlations between the dimensions. Moreover, the CFA model in which the dimension with two indicators is included produced good model fit indices.

**Table 3.** Correlation coefficients between factors.

| Factor | 1     | 2     | 3     | 4     | 5     | 6     |
|--------|-------|-------|-------|-------|-------|-------|
| 1      | 1.000 | .431  | .580  | .515  | .688  | .574  |
| 2      | .431  | 1.000 | .249  | .535  | .397  | .294  |
| 3      | .580  | .249  | 1.000 | .322  | .597  | .448  |
| 4      | .515  | .535  | .322  | 1.000 | .445  | .372  |
| 5      | .688  | .397  | .597  | .445  | 1.000 | .544  |
| 6      | .574  | .294  | .448  | .372  | .544  | 1.000 |

In Table 3, the correlation coefficients of the factors are presented. Tabachnick and Fidell (2013) suggest considering the factor correlation matrix for correlations around .32 and above. According to them, if correlations are greater than .32, then oblique rotation should be used, unless there are compelling reasons for orthogonal rotation. As it is seen from the table, most of the correlation coefficients between the factors are high, and above .32. Therefore, the oblique rotation was preferred as factor rotation method.

In Table 4 the final pattern matrix was given. The item numbers used in this matrix were according to the first form of the scale presented in the Appendix 1.

**Table 4.** Final pattern matrix.

|     | 1    | 2    | 3    | 4    | 5    | 6    |
|-----|------|------|------|------|------|------|
| M23 | .967 |      |      |      |      |      |
| M22 | .905 |      |      |      |      |      |
| M24 | .778 |      |      |      |      |      |
| M14 | .628 |      |      |      |      |      |
| M13 | .516 |      |      |      |      |      |
| M25 | .478 |      |      |      |      |      |
| M10 |      | .922 |      |      |      |      |
| M6  |      | .765 |      |      |      |      |
| M9  |      | .738 |      |      |      |      |
| M7  |      | .636 |      |      |      |      |
| M17 |      |      | .960 |      |      |      |
| M18 |      |      | .938 |      |      |      |
| M16 |      |      | .719 |      |      |      |
| M1  |      |      |      | .707 |      |      |
| M2  |      |      |      | .653 |      |      |
| M3  |      |      |      | .625 |      |      |
| M8  |      |      |      | .569 |      |      |
| M26 |      |      |      |      | .967 |      |
| M27 |      |      |      |      | .647 |      |
| M11 |      |      |      |      | .525 |      |
| M5  |      |      |      |      |      | .843 |
| M4  |      |      |      |      |      | .550 |

The items numbered as 13, 14, 22, 23, 24 and 25 loaded on Text dimension, the ones numbered 6,7,9,10 loaded on School dimension, the ones numbered 16, 17, 18 loaded on Media dimension, the ones numbered 1,2,3 and 8 loaded on Friends & Family dimension, the ones numbered 11, 26, 27 loaded on Computer dimension and lastly the ones numbered as 4 and 5 loaded on Foreigners dimension. Reliability coefficients (Cronbach's alpha) for each factor were calculated as .883, .824, .921, .786, .773 and .704 respectively.

### 3.2. Findings related to SEM

Before starting the study of structural equation modeling, data screening was also applied again and the assumptions that have to be examined in multivariate statistics were tested. Missing values were detected in answers of 14 students and these data were removed from the study. In the data, there was not a univariate outlier. Since one of the student's answers' Mahalanobis distance was higher than the related critical chi-square value, the data from this student was considered as a multivariate outlier and it was excluded from the analysis. The SEM analysis was conducted with the data of 233 students. While testing the assumptions, there was not any

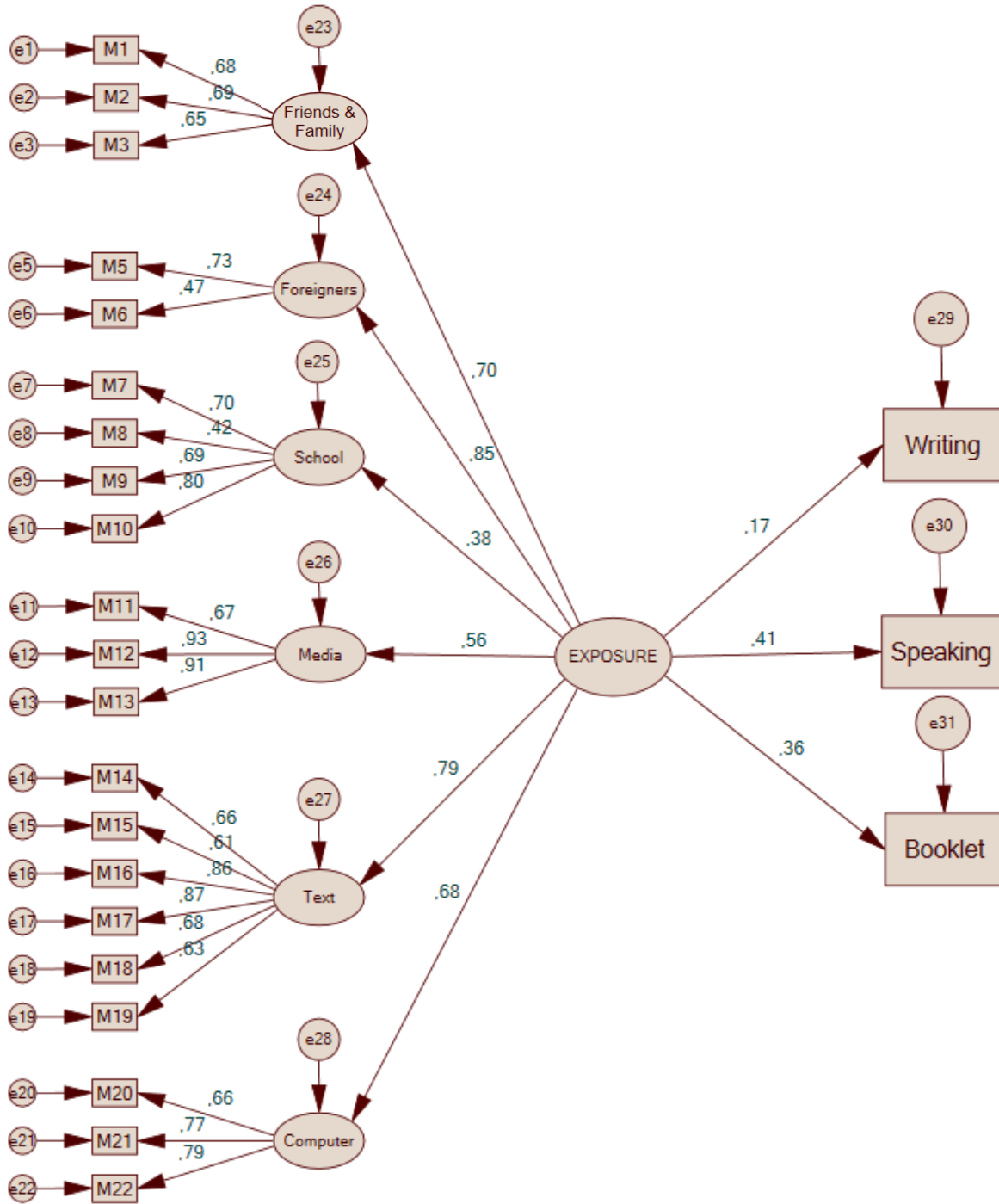
problematic variable in terms of linearity and multicollinearity, but the assumptions of univariate and multivariate normality couldn't be met. Normality tests were carried out with AMOS 23 and, in **Appendix 3**, both the univariate and multivariate normality test results are given. When we look at the skewness and kurtosis values estimated for each variable in the structural model, it is seen that there are some values indicating univariate non-normality (i.e. skewness and kurtosis values above 1). According to Brown (2015) “although univariate normality does not ensure multivariate normality, univariate non-normality does ensure multivariate non-normality” (p. 347). In other words, if our data does not meet the conditions of univariate normality, it will also not be multivariate normal data. There is, therefore, no need to investigate multivariate normality further, but Mardia's (1970) coefficient of multivariate kurtosis is also reported in the normality test results for showing another evidence of multivariate non-normality. A multivariate kurtosis value more than 10 and its critical ratio value above 1.96 together indicate a multivariate non-normal data (Byrne, 2010; Gao, Mokhtarian, & Johnston, 2008; UTEXAS, 2018). For our data, the values were estimated as 66,221 and 14,307 respectively, which shows the multivariate distribution of the data is not normal. Since the data is not normally distributed, MLR (Maximum Likelihood Estimation with Robust Standard Errors) was used as the estimation method. MLR enables conducting analysis with the data sets for which normality assumption cannot be met (Muthén & Muthén, 2010).

**Table 5.** The intervals for model fit indices and values calculated for the models.

| Model Fit Indices          | $(\chi^2 / df)$             | RMSEA                     | CFI                      | SRMR                     |
|----------------------------|-----------------------------|---------------------------|--------------------------|--------------------------|
| Good Fit                   | $0 \leq \chi^2 / df \leq 2$ | $0 \leq RMSEA \leq .05$   | $.95 \leq CFI \leq 1.00$ | $0 \leq SRMR \leq .05$   |
| Acceptable Fit             | $2 \leq \chi^2 / df \leq 5$ | $.05 \leq RMSEA \leq .10$ | $.90 \leq CFI \leq .95$  | $.05 \leq SRMR \leq .10$ |
| First Measurement Model    | 2.43                        | .078                      | .852                     | .077                     |
| Modified Measurement Model | 1.84                        | .060                      | .917                     | .069                     |
| Structural Model           | 1.69                        | .054                      | .920                     | .069                     |

Firstly, a measurement model was specified to confirm the factor structure of the data belonging to exposure to language scale which was used in the study. Since the fit indices calculated for the first measurement model did not produce acceptable values for good model fit, the modifications, which cause the most decrease in chi-square value, were carried out after examinations of the modification indices. Firstly, item 4 (My schoolmates speak English) was deleted from the model, because it cross-loaded on the dimension of School. Then the errors of item 14 (I read web pages in English) and item 15 (I follow blogs in English) were allowed to co-vary. After these modifications which were also conceptually reasonable, the measurement model produced acceptable and good model fit indices ( $\chi^2 / d.f. = 1.839$ , RMSEA = 0.060, CFI = 0.917, SRMR = 0.069) (Bollen, 1989; Byrne, 2010, 2012; Kline, 2016). The model fit indices estimated for the first measurement model, the modified measurement model, and the structural model are given in [Table 5](#).





**Figure 2.** Structural Model

After the measurement model was analyzed, a structural model was formed in which the variable of exposure was specified as the exogenous latent variable and the variables of writing, speaking and booklet were included as endogenous dependent variables. The structural model produced good and acceptable fit indices ( $\chi^2 / d.f. = 1.690$ , RMSEA = 0.054, CFI = 0.920, SRMR = 0.069) without any need for a modification. The values estimated for the parameters in the structural model are given both on the diagram in Figure 2 and in Table 6. For each value calculated for the parameters in the structural model the p values are estimated below .05 and therefore they are all statistically significant as in the measurement model.

According to the results obtained from the structural model, exposure to English has significant effects on speaking, writing and booklet scores of the students. While the variable which exposure affects most is speaking ( $\gamma = .405$ ), the one which is affected least is the writing variable ( $\gamma = .174$ ). To be able to make interpretations about the level of the effects on the dependent variables, the standardized measure of effect size ( $f^2$ ) which was suggested to use in regression-based studies was also calculated (Cohen, 1988) and they are given in Table 6 too. It is observed that the effects of exposure to English on the variances of three dependent variables are not high. While the effect of exposure on the variances of writing and the total score are small, it is medium on the variance of speaking.

**Table 6.** The parameters estimated for the dependent variables.

| Variable | Effect of Exposure | $r^2$ | $f^2$ (effect size) |
|----------|--------------------|-------|---------------------|
| Writing  | $\gamma = .174$    | .032  | .033 (small)        |
| Speaking | $\gamma = .405$    | .165  | .198 (medium )      |
| Booklet  | $\gamma = .364$    | .120  | .136 (small)        |

#### 4. DISCUSSION and CONCLUSION

In this study firstly, a scale was developed in order to measure university students' language exposure to English, which has a considerable effect in language acquisition, and secondly, the effect of the language exposure on some components of the English achievement was investigated with a structural model.

The factor structure of the scale was found as six-dimensional. That is to say, the individuals learning English are exposed to English from six different sources which are "Friends & Family", "Foreigners" (by making contact with friends and with the foreigners speaking English, "School" (in English courses or in courses taught in English), "Media" (by watching series or movies in English), "Text" (by reading books or newspapers in English), and "Computer" (by playing games or using software in English).

According to the results obtained from the structural model, exposure to English has significant effects on writing in English, speaking English and booklet score which is a total score of reading and listening comprehension, grammar and vocabulary knowledge.

In this study, only the variable of exposure to English was included as a predictor of English achievement. In future studies, by using the scale developed in this study, with more complex models, the effect of exposure on English achievement will be investigated with other factors affecting language acquisition. Moreover, these models may also include the factors affecting exposure to English.

This study contributes the existing literature in two important ways. Firstly, although exposure to a language is found to be as an important aspect of language learning, tools to measure the amount of language exposure is limited to the questionnaires (Derwing et al., 2007; Djigunovi'c et al., 2008; Olsson, 2012; Peters, 2018). The questionnaires only allow researchers determine the amount of exposure for each item, but they cannot be used to sum the language exposure, because for these studies construct validity evidence were not performed or reported. The only exception is the study of Gökcan and Çobanoğlu Aktan (2016). Nevertheless, the scale developed in that study aimed to measure the exposure to language for elementary students. Considering the age group of this study, which is university students, a new tool, which reflects the sources of exposure for this age group, was necessary. Moreover, even if some of the items and the sub-scales were similar in the scales for elementary and university students, it was

necessary to obtain an evidence for the construct validity of the new scale for university students. Secondly, this study contributes the literature in terms of providing evidence for how language exposure is related to speaking and writing language skills by structural equation modeling. This analysis method allows considering the measurement error in the proposed model.

In addition to investigating relation among speaking, writing, and exposure to a language, in future studies relations with other language skills such as reading and listening comprehension in English, as well as grammar and vocabulary knowledge can be investigated. Moreover, a recent study (Kilic, 2018) shows factor scores and the total score are so related to each other that they can be used interchangeably. The factor scores give as much information as the total score about the construct the scale measures. From this point of view, the separate effects of the Exposure to English scale's factors on English achievement can also be studied in future works.

## ORCID

Mustafa Gökcan  <https://orcid.org/0000-0002-2284-9967>

Derya Çobanoğlu Aktan  <https://orcid.org/0000-0002-8292-3815>

## 5. REFERENCES

- Alpar, R. (2011). *Çok değişkenli istatistiksel yöntemler [Multivariate statistical methods]* (4th ed.). Ankara: Detay Yayıncılık.
- Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, 17(3), 303-316.
- Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York: Routledge.
- Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.
- Brown, T. (2015). *Confirmatory Factor Analysis For Applied Research*. (2nd ed.). New York: The Guilford Press.
- Can, A. (2014). *SPSS ile bilimsel araştırma sürecinde nicel veri analizi [Statistical analysis in scientific research process by SPSS]*. Ankara: Pegem Akademi.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). NJ: Lawrence Erlbaum Associates.
- Cook, V. (2008). *Second language learning and language teaching* (4th ed.). London: Hodder Education.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. (2nd Edition). New Jersey: Lawrence Erlbaum Associates.
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10(7), 1-9.
- Çokluk, Ö., Şekercioğlu, G. & Büyüköztürk, Ş. (2014). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları. [Multivariate statistics for social sciences: SPSS and LISREL applications]* (3rd ed.) Ankara: Pegem Akademi Yayınları.
- Derwing, T. M., Munro, M. J. & Thomson R. I. (2007). A Longitudinal Study of ESL Learners' Fluency and Comprehensibility Development. *Applied Linguistics*, 29(3), 359-380.
- Djigunovi'c, J. M., Nikolov, M. & Otto, I. (2008). A comparative study of Croatian and Hungarian EFL students. *Language Teaching Research*, 12(3), 433-452.
- Ellis, R. (2015). *Understanding Second Language Acquisition*. Oxford: Oxford University Press.

- Gao, S., Mokhtarian, P. L., & Johnston, R. A. (2008). Nonnormality of data in structural equation models. *Transportation Research Record: Journal of the Transportation Research Board*, 2082, 116–124.
- Gökcan, M., & Çobanoğlu Aktan, D. (2016). İngilizceye Maruz Kalma Ölçeğinin Geçerlik Ve Güvenirliğinin İncelenmesi [Investigation of the validity and reliability of exposure to English scale]. In N. Akpınar Dellal & H. Yokuş (Eds), *Proceedings of international contemporary educational research congress* (pp. 283-294). Ankara: Pegem Akademi.
- Gökcan, M. & Çobanoğlu Aktan, D. (2018). Investigation of the variables related to TEOG English achievement using Language Acquisition Theory of Krashen. *Pegem Eğitim ve Öğretim Dergisi*, 8(3), 531-566.
- Hancock, G. R., & Schoonen, R. (2015). Structural equation modeling: Possibilities for language learning researchers. *Language Learning*, 65(S1), 160–184.
- Harmer, J. (2007). *The practice of English language teaching* (4th ed.). Harlow: Pearson Education.
- Kilic, A. (2018). Can Factor Scores be Used Instead of Total Score and Ability Estimation?, *International Journal of Assessment Tools in Education*, 6(1), 25-35.
- Kline, R. B. (2016), *Principles and practice of structural equation modeling*. (4th ed.). New York: The Guilford Press.
- Krashen, S. (1982). *Principles and practice in second language acquisition*. New York: Pergamon Press.
- Krashen, S. (2009). The comprehension hypothesis extended. In T. Piske & M. Young-Scholten (Eds), *Input Matters in SLA* (pp. 81-94). Bristol: Multilingual Matters.
- Lightbown, P. M. & Spada, N. (2006). *How languages are learned* (3rd ed.). Oxford: Oxford University Press.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519–530.
- Mitchell, R., Myles, F., & Marsden, E. (2013). *Second language learning theories* (3rd ed.). London: Routledge.
- Muthén, L. & Muthén, B. (2010) *Mplus User's Guide*. (6th ed.). Los Angeles, CA: Muthén & Muthén.
- Olsson, E. (2012). "Everything I read on the Internet is in English:" *On the impact of extramural English on Swedish 16-year-old pupils' writing proficiency*. Licentiate thesis. Gothenburg: Gothenburg University.
- Peters E. (2018). The effect of out-of-class exposure to English language media on learners' vocabulary knowledge. *ITL - International Journal of Applied Linguistics*, 169(1), 142-168.
- Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics*. (5th ed.). USA: Pearson Education, Inc.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
- UTEXAS (2018). Software faqs. Retrieved July 21, 2018, from <https://stat.utexas.edu/software-faqs/amos>
- Winke, P. (2014). Testing Hypotheses about Language Learning Using Structural Equation Modeling. *Annual Review of Applied Linguistics*, 34, 102-122.
- Wolf, S. D., Smit, N. & Lowie, W. (2017). Influences of early English language teaching on oral fluency. *ELT Journal*, 71(3), 341-353.
- Worthington, R. L., & Whittaker, T. A. (2006). Scale Development Research: A Content Analysis and Recommendations for Best Practices, *The Counseling Psychologist*, 34(6), 806-838.

## Appendix 1. First form of the scale

| Aşağıda verilen durumların ne sıklıkla olduğunu, size en uygun olan ifadeyi gösteren rakamı yuvarlak içine alarak belirtiniz | Hiçbir Zaman | Nadiren | Bazen | Genellikle | Her Zaman |
|--|--------------|---------|-------|------------|-----------|
| 1. Arkadaşlarım sınıf dışında İngilizce konuşur.   | 1            | 2       | 3     | 4          | 5         |
| 2. İngilizcenin konuşulduğu ortamlarda bulunurum.  | 1            | 2       | 3     | 4          | 5         |
| 3. Evimizde İngilizce konuşulur.   | 1            | 2       | 3     | 4          | 5         |
| 4. Yabancı turistlerle İngilizce konuşurum.  | 1            | 2       | 3     | 4          | 5         |
| 5. Yurt dışına seyahat ederim.   | 1            | 2       | 3     | 4          | 5         |
| 6. Öğretmenlerim İngilizce konuşur.  | 1            | 2       | 3     | 4          | 5         |
| 7. Okulumda İngilizce aktiviteler yapılır.   | 1            | 2       | 3     | 4          | 5         |
| 8. Sınıf arkadaşlarım İngilizce konuşur.   | 1            | 2       | 3     | 4          | 5         |
| 9. Okulum İngilizce konuşmamızı teşvik eder.   | 1            | 2       | 3     | 4          | 5         |
| 10. Okulda dersler İngilizce işlenir.  | 1            | 2       | 3     | 4          | 5         |
| 11. İnternet ortamında İngilizce sohbet ederim.  | 1            | 2       | 3     | 4          | 5         |
| 12. İngilizce mesajlaşırım (e-mail, sms, whatsapp)   | 1            | 2       | 3     | 4          | 5         |
| 13. İnternette İngilizce web sayfalarını okurum.   | 1            | 2       | 3     | 4          | 5         |
| 14. İnternette İngilizce blogları takip ederim.  | 1            | 2       | 3     | 4          | 5         |
| 15. İngilizce sosyal medya sayfalarını takip ederim  | 1            | 2       | 3     | 4          | 5         |
| 16. İngilizce şarkı dinlerim.  | 1            | 2       | 3     | 4          | 5         |
| 17. İngilizce dizi izlerim.  | 1            | 2       | 3     | 4          | 5         |
| 18. İngilizce film izlerim.  | 1            | 2       | 3     | 4          | 5         |
| 19. İngilizce çizgi film- anime izlerim.   | 1            | 2       | 3     | 4          | 5         |
| 20. İngilizce televizyon programı izlerim.   | 1            | 2       | 3     | 4          | 5         |
| 21. İngilizce youtube videoları izlerim.   | 1            | 2       | 3     | 4          | 5         |
| 22. İngilizce dergi okurum.  | 1            | 2       | 3     | 4          | 5         |
| 23. İngilizce gazete okurum.   | 1            | 2       | 3     | 4          | 5         |
| 24. İngilizce kitap okurum.  | 1            | 2       | 3     | 4          | 5         |
| 25. İngilizce karikatür okurum.  | 1            | 2       | 3     | 4          | 5         |
| 26. İngilizce bilgisayar oyunu oynarım.  | 1            | 2       | 3     | 4          | 5         |
| 27. İngilizce bilgisayar programı kullanırım.  | 1            | 2       | 3     | 4          | 5         |

**Appendix 2.** Last form of the scale

|    | Aşağıda verilen durumların ne sıklıkla olduğunu, size en uygun olan ifadeyi gösteren rakamı yuvarlak içine alarak belirtiniz | Hiçbir Zaman | Nadiren | Bazen | Genellikle | Her Zaman |
|----|--|--------------|---------|-------|------------|-----------|
| 1  | Arkadaşlarım sınıf dışında İngilizce konuşur.  | 1            | 2       | 3     | 4          | 5         |
| 2  | İngilizcenin konuşulduğu ortamlarda bulunurum.   | 1            | 2       | 3     | 4          | 5         |
| 3  | Evimizde İngilizce konuşulur.  | 1            | 2       | 3     | 4          | 5         |
| 4  | Sınıf arkadaşlarım İngilizce konuşur.  | 1            | 2       | 3     | 4          | 5         |
| 5  | Yabancı turistlerle İngilizce konuşurum.   | 1            | 2       | 3     | 4          | 5         |
| 6  | Yurt dışına seyahat ederim.  | 1            | 2       | 3     | 4          | 5         |
| 7  | Öğretmenlerim İngilizce konuşur.   | 1            | 2       | 3     | 4          | 5         |
| 8  | Okulumda İngilizce aktiviteler yapılır.  | 1            | 2       | 3     | 4          | 5         |
| 9  | Okulum İngilizce konuşmamızı teşvik eder.  | 1            | 2       | 3     | 4          | 5         |
| 10 | Okulda dersler İngilizce işlenir.  | 1            | 2       | 3     | 4          | 5         |
| 11 | İngilizce şarkı dinlerim.  | 1            | 2       | 3     | 4          | 5         |
| 12 | İngilizce dizi izlerim.  | 1            | 2       | 3     | 4          | 5         |
| 13 | İngilizce film izlerim.  | 1            | 2       | 3     | 4          | 5         |
| 14 | İnternette İngilizce web sayfalarını okurum.   | 1            | 2       | 3     | 4          | 5         |
| 15 | İnternette İngilizce blogları takip ederim.  | 1            | 2       | 3     | 4          | 5         |
| 16 | İngilizce dergi okurum.  | 1            | 2       | 3     | 4          | 5         |
| 17 | İngilizce gazete okurum.   | 1            | 2       | 3     | 4          | 5         |
| 18 | İngilizce kitap okurum.  | 1            | 2       | 3     | 4          | 5         |
| 19 | İngilizce karikatür okurum.  | 1            | 2       | 3     | 4          | 5         |
| 20 | İnternet ortamında İngilizce sohbet ederim.  | 1            | 2       | 3     | 4          | 5         |
| 21 | İngilizce bilgisayar oyunu oynarım.  | 1            | 2       | 3     | 4          | 5         |
| 22 | İngilizce bilgisayar programı kullanırım.  | 1            | 2       | 3     | 4          | 5         |



---

**Appendix 3. Normality Test Results**

Assessment of normality:

| <b>Variable</b>     | <b>min</b> | <b>max</b> | <b>skew</b> | <b>c.r.</b> | <b>kurtosis</b> | <b>c.r.</b> |
|---------------------|------------|------------|-------------|-------------|-----------------|-------------|
| <b>B</b>            | 20.000     | 91.000     | -.223       | -1.391      | -.480           | -1.495      |
| <b>S</b>            | 25.000     | 100.000    | -.238       | -1.486      | -.458           | -1.428      |
| <b>W</b>            | 25.000     | 100.000    | -.314       | -1.959      | -.343           | -1.068      |
| <b>M22</b>          | 1.000      | 5.000      | -.238       | -1.483      | -1.288          | -4.013      |
| <b>M19</b>          | 1.000      | 5.000      | .697        | 4.343       | -.702           | -2.187      |
| <b>M18</b>          | 1.000      | 5.000      | .593        | 3.694       | -.353           | -1.100      |
| <b>M17</b>          | 1.000      | 5.000      | 1.366       | 8.513       | 1.199           | 3.736       |
| <b>M16</b>          | 1.000      | 5.000      | .889        | 5.542       | -.042           | -.132       |
| <b>M13</b>          | 1.000      | 5.000      | -1.082      | -6.742      | .232            | .722        |
| <b>M10</b>          | 1.000      | 5.000      | -1.170      | -7.293      | 1.029           | 3.208       |
| <b>M9</b>           | 1.000      | 5.000      | -.554       | -3.453      | -.451           | -1.406      |
| <b>M3</b>           | 1.000      | 5.000      | 1.765       | 11.002      | 3.126           | 9.740       |
| <b>M20</b>          | 1.000      | 5.000      | .339        | 2.114       | -.791           | -2.464      |
| <b>M21</b>          | 1.000      | 5.000      | -.229       | -1.427      | -1.446          | -4.504      |
| <b>M14</b>          | 1.000      | 5.000      | .166        | 1.031       | -.833           | -2.597      |
| <b>M15</b>          | 1.000      | 5.000      | .410        | 2.553       | -.832           | -2.592      |
| <b>M11</b>          | 1.000      | 5.000      | -.940       | -5.859      | .214            | .667        |
| <b>M12</b>          | 1.000      | 5.000      | -1.084      | -6.752      | .280            | .871        |
| <b>M7</b>           | 1.000      | 5.000      | -1.261      | -7.859      | 1.884           | 5.871       |
| <b>M8</b>           | 1.000      | 5.000      | -.530       | -3.303      | -.559           | -1.743      |
| <b>M5</b>           | 1.000      | 5.000      | .428        | 2.670       | -.705           | -2.195      |
| <b>M6</b>           | 1.000      | 5.000      | 1.701       | 10.599      | 1.918           | 5.975       |
| <b>M1</b>           | 1.000      | 5.000      | 1.000       | 6.234       | .780            | 2.431       |
| <b>M2</b>           | 1.000      | 5.000      | .346        | 2.156       | -.324           | -1.010      |
| <b>Multivariate</b> |            |            |             |             | 66.221          | 14.307      |

## Adaptation of Physics Metacognition Inventory to Turkish

Zeynep Koyunlu Ünlü <sup>1,\*</sup> İlbilge Dökme <sup>2</sup>

<sup>1</sup> Yozgat Bozok University, Faculty of Education, Department of Primary Education, Yozgat, Turkey

<sup>2</sup> Gazi University, Faculty of Gazi Education, Department of Mathematics and Science Education, Ankara, Turkey

### ARTICLE HISTORY

Received: 23 November 2018

Revised: 23 February 2019

Accepted: 05 March 2019

### KEYWORDS

Physics education,  
Metacognition,  
Scale adaptation

**Abstract:** This study aimed to adapt the Physical Metacognition Inventory (PMI) developed by Taasobshirazi and Farley (2013) to Turkish. PMI consists of 24 items and six factors. The scale items were translated into Turkish by the researchers, and a Turkish-English comprehensibility form was prepared to elicit the opinions of Turkish-English language experts. After making the necessary revision according to the feedback of the experts, a confirmatory factor analysis (CFA) was undertaken. A total of 554 students participated in the research, selected from prospective teachers enrolled in the science teaching and classroom teaching programs offered by education faculties or prospective engineers studying in engineering faculties. The results of CFA revealed that the factors and related items of the adapted scale were the same as in the original version. The reliability of measurement was calculated as 0.93 for the whole scale. The adapted PMI presented in this research can be applied to evaluate the level of metacognition used by high school and university students in solving physics problems.

## 1. INTRODUCTION

Metacognition refers to knowledge and cognition about a cognitive phenomenon (Flavell, 1979). Thinking about metacognition is to become aware of what we know and what we do not know (Blakey & Spence, 1990; Lai, 2011). In other words, it means reflecting, understanding and managing one's learning (Schraw & Dennison, 1994). When the definitions related to metacognition are examined, it is observed that they generally focus on the individual's awareness and control of his/her knowledge and processes related to learning while cognition is more related to the mental learning of individuals.

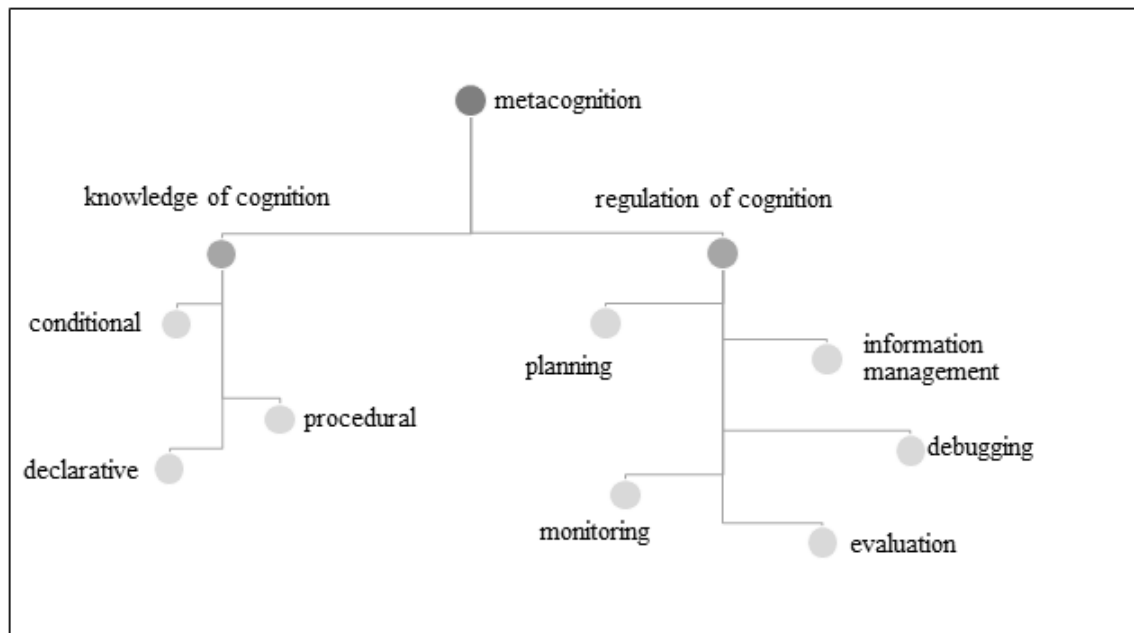
Metacognition consists of two dimensions: knowledge of cognition related to one's own cognitive resources and regulation of cognition containing information used in the problem-solving process. Knowledge of cognition comprises the sub-scales of declarative, procedural and conditional knowledge, and regulation of cognition encompasses planning, monitoring,

---

CONTACT: Zeynep Koyunlu Ünlü ✉ [zeynepko.unlu@gmail.com](mailto:zeynepko.unlu@gmail.com) 📍 Yozgat Bozok University, Faculty of Education, Division of Classroom Instruction Education, Yozgat, Turkey

ISSN-e: 2148-7456 / © IJATE 2019

evaluation, debugging, and information management (Taasoobshirazi & Farley, 2013). The subdimensions of the metacognition are presented in Figure 1.



**Figure 1.** Sub-scales of metacognition (Taasoobshirazi & Farley, 2013)

The knowledge of cognition as the first component of metacognition during problem solving tasks refers to the effect of the students' performance in relation to how they use these strategies appropriately in accordance with the task (Brown, 1978). Declarative knowledge, a sub-scale of knowledge of cognition, is related to the factors affecting the person himself/herself and his/her learning performance. Procedural knowledge concerns knowing what strategy to use and when. Conditional knowledge is knowing when and why to use the remaining components of knowledge of cognition (Taasoobshirazi & Farley, 2013). The regulation of cognition as the second component of metacognition also refers to how learners monitor, control, and regulate their cognition and learning (Pintrich, 2002; Schraw, Crippen & Hartley, 2006; Schraw & Moshman, 1995). Of the sub-scales of regulation of cognition, planning concerns goal setting, activation of past information, and arranging time; monitoring is the self-evaluation of an individual at certain intervals; evaluation refers to reviewing one's learning and associated products and process; debugging is the elimination of unnecessary information; and lastly information management concerns using individual-specific strategies to solve a problem effectively (Taasoobshirazi & Farley, 2013).

Over the past decades, metacognition received much attention in the science education literature. Particularly in recent years, metacognitive instruction has been shown to improve students' conceptual understanding of science (Abd-El-Khalick & Akerson 2009; Colthorpe, Sharifirad, Ainscough, Anderson & Zimbardi, 2018) and develop their higher-order thinking (Ghanizadeh, 2018) and problem-solving skills (Akben, 2018). Also, metacognition has been considered as one of the most important issues in the students' success in problem-solving. For example, in physics classes, students should practice meta-cognitively throughout processes of solving physics problems by defining the goals in the problem, mental representation of the problem, selecting the proper strategies, connecting prior knowledge, planning, monitoring, and evaluation of possible solutions (Güss & Wiley, 2007; Taasoobshirazi & Farley, 2013; Taasoobshirazi, Bailey & Farley, 2015).

Research has revealed that some students are unable to solve non-routine physics problems

because of the lack of the metacognitive skills or awareness (Selçuk, Çalışkan & Erol 2007; Anzai & Yokoyama 1984; Stewart & Rudolph 2001). Students can acquire metacognitive knowledge theoretically, but there is a strong need for practical implementation during problem-solving processes in physics lessons (Georghiades, 2004; Thomas, 2012; Zohara & Barzilai, 2013; Hutner & Markman, 2016). Metacognitive tasks in physics courses allow the students to gain experience and develop metacognitive skills (Veenman & Spaans 2005; Veenman, 2011).

The first attempt to include metacognitive thinking in the physics problem-solving process seems to be to reveal the metacognitive awareness of students. Because it is important to determine what level the students have before developing their metacognition (Öztürk, 2017). At this point, one more important element is the development or adaptation of standardized instruments for the participants. Therefore, the first aim of this study is to adapt into Turkish the Physics Metacognition Inventory (PMI) developed by Taasoobshirazi and Farley, (2013), which is a 5-point, Likert-type scale. However, first it is necessary to determine the cognitive status of students prior to commencing the research. One potential reason for the lack of studies examining the role of metacognition on physics problem-solving is the absence of an inventory that measures metacognition for science problem-solving. Most of the existing research examining metacognition for problem-solving in science has done so using primarily verbal interviews or a small set of researcher-developed items (Rozenchwajg, 2003), and this indicates that students who are more metacognitive during physics problem-solving are more likely to correctly solve the problems (Neto & Valente, 1997; Rozenchwajg, 2003). The lack of research on the role of metacognition in physics problem-solving is problematic given the significance of problem-solving for success and improvement in physics (Chi, 2006). A review of the literature revealed the availability of attitude scales related to physics teaching and physics laboratories (Kurnaz & Yiğit, 2010; Nuhoglu & Yalçın, 2004; Tekbıyık & Akdeniz, 2010); however, there is no measurement tool in the Turkish literature for measuring physics metacognition. In the current study, PMI developed by Taasoobshirazi and Farley (2013) was adapted to Turkish to fill the gap in the national literature and guide researchers and practitioners.

## **2. METHOD**

This was a scale adaptation study. Scale adaptation refers to the process in which a scale that was developed in another language and proven to be reliable and valid is adapted to another language and culture and made ready for use through reliability and validity tests (Seçer, 2015). In this research, PMI developed by Taasoobshirazi and Farley (2013) was adapted to Turkish.

### **2.1. Study Group**

The study group was selected according to the criterion sampling technique, in which observation units may be persons, objects or situations with specific characteristics (Patton, 2002). The criterion in this study was determined as the participants in this research group taking the physics course at the university level.

The study group consisted of prospective teachers enrolled in the science teaching and classroom teaching programs and prospective engineers studying in the faculty of technology in two universities located in the Central Anatolia Region of Turkey. [Table 1](#) presents detailed information about the main study group.

A total of 96 students (62 female, 34 male) participated in the pilot study and 458 students (347 female, 111 male) in the main study. In addition, one Turkish language expert, one English language expert, and six field experts with a PhD in science and physics education

were consulted during the scale adaptation process. Analyzes were carried out from the data of 458 students who participated in main study. The pilot study focused on whether there were any problems understanding of PMI.

**Table 1.** Frequency and percentages of the participant prospective teachers according to gender and department

| Variable   | Sub-variable       | Frequency | Percentage |
|------------|--------------------|-----------|------------|
| Gender     | Female             | 347       | 78         |
|            | Male               | 111       | 22         |
| Department | Science Teaching   | 223       | 49         |
|            | Classroom Teaching | 184       | 40         |
|            | Engineering        | 51        | 11         |

## 2.2. PMI

The PMI instrument developed by Taasobshirazi and Farley (2013) is based on the theory of processing information. It consists of 24 items presented under the two main dimensions of knowledge of cognition and regulation of knowledge, which have a total of six sub-scales: declarative, procedural and conditional (sub-scales of the former), and planning, monitoring, evaluation, debugging, and information management (sub-scales of the latter). [Table 2](#) presents the items under the PMI factors.

**Table 2.** Items included in the original PMI

| Factors  | Items               |
|--|---------------------|
| Knowledge of cognition: declarative, procedural, conditional | 5, 6, 7, 11, 12, 13 |
| Regulation of knowledge: information management              | 4, 10, 18, 23       |
| monitoring   | 2, 15, 16, 21       |
| evaluation   | 8, 9, 17            |
| debugging  | 3, 22               |
| planning   | 1, 14, 19, 20, 24   |

The number of items under the factors of PMI varies between two and six ([Table 2](#)). None of the items contain a negative statement. This inventory is based on a 5-point Likert-type scale with the following possible responses: always true of myself (5), almost always true of myself (4), sometimes true of myself (3), rarely true of myself (2), and never true of myself (1) (Taasobshirazi & Farley, 2013).

## 2.3. Procedure

As a matter of academic courtesy, the corresponding author of the PMI study (Gita Taasobshirazi) was contacted and permission was obtained to adapt the scale into Turkish. The scale was first translated into Turkish by the researchers. At this stage, one Turkish and one English language expert were consulted. In the first stage, the translated version of the inventory was completed by five prospective teachers to confirm that the items were understandable. Then, a Turkish-English comprehensibility form was prepared to elicit the opinion of six field experts from science and physics teaching. The correlation coefficients of each expert's score was calculated and the necessary corrections were undertaken by the researchers. After revision according to the feedback from the experts, the Turkish version of the scale was administered to 96 students in a pilot study. The analysis of data obtained from the pilot study revealed that the factors and item distribution were in line with the original scale. Thus, the main study was undertaken with 458 participants and data analysis was conducted.

## 2.4. Data Analysis

AMOS 21 and SPSS 21 programs were used for data analysis. A confirmatory factor analysis (CFA) was performed using the AMOS 21 program. In CFA, a previously determined model or hypothesis on the relationship between variables is tested (Büyüköztürk, 2004). In this study, the first level multi-factor model for the adapted PMI was tested. The variables that can be observed in this model are grouped under more than one independent dimension (Meydan & Şeşen, 2015). Using the SPSS 21 program, Cronbach's alpha values were calculated for the whole PMI and each factor to determine reliability of measurement, the corrected item-total correlations of the factors and t-test between the upper and lower 27% scores were performed, the mean and standard deviation values and the correlations between the sub-scales were determined, and the test-retest reliability was undertaken. In addition, for criterion validity PMI scores were analyzed according to the participants' gender and department.

## 3. RESULTS

### 3.1. Translation of PMI into Turkish

After the translation of PMI into Turkish by the authors, one Turkish and one English language expert were consulted. The Turkish-English comprehensibility form was completed by six science and physics teachers. The experts scored the translation of each item from 1 to 5. The correlation coefficient between the scores of the experts and the mean item scores were calculated. In this process, it was checked whether the mean score given to the translation of the items in the scale was 4.0 or above, and the standard deviation was 0.7 or below. The mean score was calculated as 4.3 and the standard deviation as 0.4. Language experts have suggested some words to be changed. In addition, the recommendations made by the experts do not contain a substance that does not comply with the Turkish culture. This is due to the lack of direct translation.

### 3.2. Results of Reliability Analysis

The measurement reliability values of the Turkish version of the whole PMI and its factors were calculated using the SPSS 21 program. [Table 3](#) shows the measurement reliability values of the sub-scales of both the original and adapted versions of PMI.

**Table 3.** *The measurement reliability values of the sub-scales included in the English and Turkish versions of PMI*

| Sub-scales | English version | Turkish version |
|------------|-----------------|-----------------|
| Factor 1   | 0.90            | 0.87            |
| Factor 2   | 0.91            | 0.86            |
| Factor 3   | 0.87            | 0.8             |
| Factor 4   | 0.78            | 0.8             |
| Factor 5   | 0.92            | 0.72            |
| Factor 6   | 0.68            | 0.74            |

The measurement reliability of the sub-scales in the English and Turkish PMI ranged from 0.92 to 0.68, and 0.87 to 0.72, respectively. This suggests that the measurement reliability of the PMI sub-scales of the adapted scale was at an acceptable level (Nunally, 1978). The overall measurement reliability of the Turkish PMI was calculated as .93.

### 3.3. Results of CFA

The original six-factor structure of PMI was tested via CFA using the AMOS 21 program. Fit indices, namely chi-square goodness ( $\chi^2$ ), goodness-of-fit index (GFI), adjusted GFI (AGFI), and the root mean square error of approximation (RMSEA) were examined. A scale is



considered to be acceptable if the values for these indices are as follows: less than 5 for the ratio of  $\chi^2$  to the degree of freedom, greater than .90 for GFI, greater than .80 for AGFI, greater than .90 for CFI, and .05 to 0.8 for RMSEA (Klein, 1998). All the fit indices obtained were within acceptable psychometric ranges ( $\chi^2/df=3.55$ , GFI=0.86, AGFI=0.82, CFI=0.89, RMSEA=0.07). The tested model is shown in Figure 2.

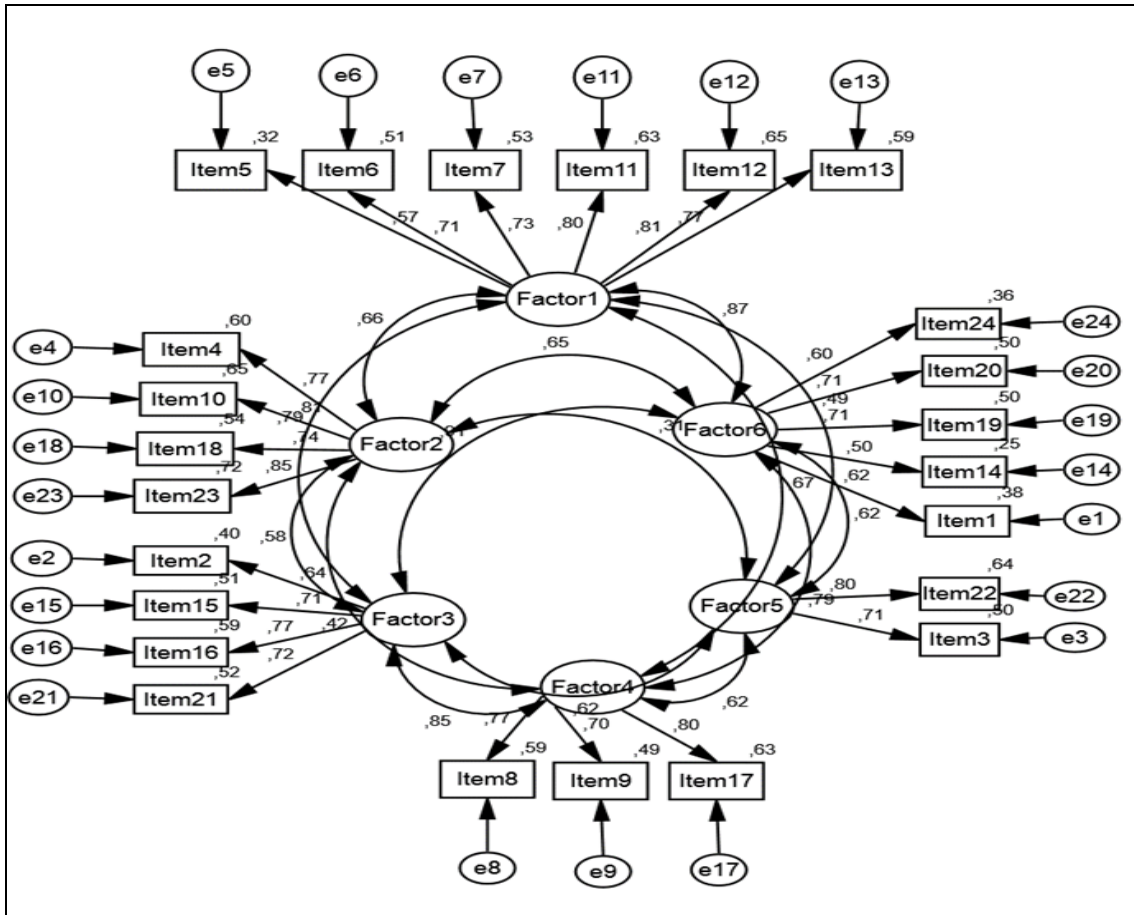


Figure1. The results of CFA of the tested model, N = 458,  $\chi^2/df = 3.55$ ,  $p < 0.001$

The factor load values of the items in the scale varied between 0.49 and 0.84 (Figure 2). The ranges of the load values were 0.56-0.8 for Factor 1, 0.73-0.84 for Factor 2, 0.63-0.76 for Factor 3, 0.7-0.79 for Factor 4, 0.7-0.8 for Factor 5, and 0.49-0.76 for Factor 6. According to these results, all values were statistically significant ( $p < 0.001$ ).

### 3.4. Results of PMI Item Analysis

In order to determine the discriminatory levels of the items in the adapted PMI and their predictive power for the total score, the corrected item-total correlation was calculated using Pearson's product moment correlation, and upper and lower 27% group comparisons were undertaken employing an independent samples t-test. Table 4 presents the results of the t-test conducted to determine the significance of the differences between the item mean scores of the upper 27% and lower 27% groups created according to the PMI total scores.

The corrected item-total correlations ranged from 0.45 to 0.71, and the t values were significant ( $p < .05$ ) (Table 4). Item-total correlation coefficients of  $r \geq .40$  are classified as very good (Nunnally & Bernstein, 1994). In this context, for the adapted inventory, the correlation between the items and the total scale was very good ( $r \geq .40$ ). The significance of the t values for the differences between the lower and upper groups was considered to be evidence for the

discriminative power of the items (Erkuş, 2012; Tezbaşaran, 1996). According to these criteria, it can be stated that all the items in the scale were discriminative.

**Table 4.** The corrected item-total correlations of PMI factors and the t-test results of the comparison between the upper 27% and lower 27% groups

| Factors  | Item No | Corrected Item-Total Correlation | t      |
|--|---------|----------------------------------|--------|
| Factor 1: Knowledge of cognition:<br>declarative,<br>conditional | 5       | 0.6                              | 12.75* |
|  | 6       | 0.66                             | 14.12* |
|  | 7       | 0.69                             | 14.71* |
|  | 11      | 0.69                             | 16.1*  |
|  | 12      | 0.71                             | 17.05* |
|  | 13      | 0.68                             | 16.43* |
| Factor 2: Regulation of cognition:<br>information management     | 4       | 0.55                             | 11.87* |
|  | 10      | 0.57                             | 12.02* |
|  | 18      | 0.6                              | 14.15* |
|  | 23      | 0.66                             | 15.02* |
| Factor 3: Regulation of cognition:<br>monitoring                 | 2       | 0.62                             | 12.04* |
|  | 15      | 0.65                             | 13.8*  |
|  | 16      | 0.69                             | 15.82* |
|  | 21      | 0.66                             | 14.77* |
| Factor 4: Regulation of cognition:<br>evaluation                 | 8       | 0.63                             | 14.52* |
|  | 9       | 0.53                             | 11.84* |
|  | 17      | 0.67                             | 15.4*  |
| Factor 5: Regulation of cognition:<br>debugging                  | 3       | 0.45                             | 8.77*  |
|  | 22      | 0.52                             | 10.65* |
| Factor 6: Regulation of cognition:<br>planning                   | 1       | 0.61                             | 11.13* |
|  | 14      | 0.52                             | 11.13* |
|  | 19      | 0.67                             | 13.8*  |
|  | 20      | 0.67                             | 14.46* |
|  | 24      | 0.58                             | 11.89* |

### 3.5. Correlation between the Sub-Scales of PMI

Correlation values between the sub-scales of PMI can be seen in Table 5. The correlation values between the sub-scales of PMI ranged from 0.24 to 0.88 (Table 5). In addition, when these results were examined together with the correlation coefficients given in Table 4, it was observed that the values generally indicated a moderate and high level of relationship between the sub-scales. (Büyüköztürk, 2014).

**Table 5.** Correlation values between the sub-scales of PMI

|          | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 | Factor 6 | Total  |
|----------|----------|----------|----------|----------|----------|----------|--------|
| Factor 1 | 1        | 0.57**   | 0.67**   | 0.56**   | 0.4**    | 0.75**   | 0.88** |
| Factor 2 | 0.57**   | 1        | 0.49**   | 0.34**   | 0.24**   | 0.55**   | 0.72** |
| Factor 3 | 0.67**   | 0.49**   | 1        | 0.66**   | 0.47**   | 0.73**   | 0.85** |
| Factor 4 | 0.56**   | 0.34**   | 0.66**   | 1        | 0.47**   | 0.6**    | 0.74** |
| Factor 5 | 0.4**    | 0.24**   | 0.47**   | 0.47**   | 1        | 0.45**   | 0.56** |
| Factor 6 | 0.75**   | 0.55**   | 0.73**   | 0.6**    | 0.45**   | 1        | 0.88** |
| Total    | 0.88**   | 0.72**   | 0.85**   | 0.74**   | 0.56**   | 0.88**   | 1      |

### 3.6. Findings about Criterion Relation Validity

For criterion validity PMI scores were analyzed according to the participants' gender and department. The results of t-test whether there was a significant difference between female and male students' PMI scores, are shown in Table 6.

**Table 6.** *t*-test results for participants' PMI Scores according to the gender

| PMI and factors  | Groups | M     | S     | t    | p     |
|--|--------|-------|-------|------|-------|
| Factor 1: Knowledge of cognition: declarative, procedural, conditional | Female | 20.6  | 4.57  | 0.48 | 0.62  |
|  | Male   | 20.3  | 4.39  |      |       |
| Factor 2: Regulation of cognition: information management              | Female | 12.76 | 3.74  | 1.26 | 0.2   |
|  | Male   | 12.25 | 3.67  |      |       |
| Factor 3: Regulation of cognition: monitoring                          | Female | 14.89 | 3     | 0.83 | 0.4   |
|  | Male   | 14.63 | 2.73  |      |       |
| Factor 4: Regulation of cognition: evaluation                          | Female | 11.67 | 2.45  | 1.01 | 0.31  |
|  | Male   | 11.4  | 2.38  |      |       |
| Factor 5: Regulation of cognition: debugging                           | Female | 8.19  | 1.6   | 2.32 | 0.02* |
|  | Male   | 7.77  | 1.79  |      |       |
| Factor 6: Regulation of cognition: planning                            | Female | 18.21 | 3.53  | 0.4  | 0.68  |
|  | Male   | 18.36 | 3.36  |      |       |
| Total PMI  | Female | 86.36 | 15.27 | 0.94 | 0.34  |
|  | Male   | 84.8  | 14.38 |      |       |

According to Table 6 students' scores for physics metacognition didn't differ significantly factor 1 ( $t_{(456)}=0.48$ ,  $p>.05$ ), factor 2 ( $t_{(456)}=1.26$ ,  $p>.05$ ), factor 3 ( $t_{(456)}=0.83$ ,  $p>.05$ ), factor 4 ( $t_{(456)}=1.01$ ,  $p>.05$ ), factor 6 ( $t_{(456)}=0.4$ ,  $p>.05$ ) and total PMI ( $t_{(456)}=0.94$ ,  $p>.05$ ). However students' scores for factor 5 differed in favor of the female students ( $t_{(456)}=2.32$ ,  $p<.05$ ). For determining if there was a significant difference students' PMI scores and departments ANOVA test was used. Table 7 shows the results of ANOVA test for the participants' PMI scores according to the department.

**Table 7.** ANOVA results for participants' PMI scores according to department

| PMI and factors  | Groups       | M     | S    | F    | p     | Post-hoc   |
|--|--------------|-------|------|------|-------|------------|
| Factor 1: Knowledge of cognition: declarative, procedural, conditional | Classroom t. | 19.96 | 5.13 | 2.73 | 0.06  | -          |
|  | Science t.   | 21    | 3.96 |      |       |            |
|  | Engineering  | 20    | 4.3  |      |       |            |
| Factor 2: Regulation of cognition: information management              | Classroom t. | 12.52 | 3.59 | 0.15 | 0.85  | -          |
|  | Science t.   | 12.72 | 3.7  |      |       |            |
|  | Engineering  | 12.73 | 4.45 |      |       |            |
| Factor 3: Regulation of cognition: monitoring                          | Classroom t. | 14.48 | 3.31 | 6.21 | 0.00* | S>C<br>S>E |
|  | Science t.   | 15.28 | 2.5  |      |       |            |
|  | Engineering  | 13.92 | 2.99 |      |       |            |
| Factor 4: Regulation of cognition: evaluation                          | Classroom t. | 11.22 | 2.73 | 4.45 | 0.01* | S>C        |
|  | Science t.   | 11.93 | 2.16 |      |       |            |
|  | Engineering  | 11.5  | 2.23 |      |       |            |
| Factor 5: Regulation of cognition: debugging                           | Classroom t. | 8     | 1.74 | 6.03 | 0.00* | S>E        |
|  | Science t.   | 8.29  | 1.45 |      |       |            |
|  | Engineering  | 7.38  | 2    |      |       |            |
| Factor 6: Regulation of cognition: planning                            | Classroom t. | 17.68 | 3.88 | 4.75 | 0.00* | S>C        |
|  | Science t.   | 18.73 | 3.15 |      |       |            |
|  | Engineering  | 18    | 3    |      |       |            |
| Total PMI  | Classroom t. | 83.89 | 17.4 | 4.14 | 0.01* | S>C        |
|  | Science t.   | 87.98 | 12.8 |      |       |            |
|  | Engineering  | 84.3  | 14   |      |       |            |

According to Table 7 students' scores for physics metacognition didn't differ significantly regarding to the department in factor 1 ( $F_{(2, 455)}=2.73$ ,  $p>.05$ ) and, factor 2 ( $F_{(2, 455)}=0.15$ ,  $p>.05$ ). There is a significant difference between students who attend science and classroom

teaching program in factor 3 ( $F_{(2, 455)}=6.21, p<.05$ ), factor 4 ( $F_{(2, 455)}=4.45, p<.05$ ), factor 6 ( $F_{(2, 455)}=4.75, p<.05$ ) and total PMI scores ( $F_{(2, 455)}=4.14, p<.05$ ) in favor of science teaching students. Also there is a significant difference between students who attend science classroom teaching and engineering program in factor 3 ( $F_{(2, 455)}=6.21, p<.05$ ) and factor 5 ( $F_{(2, 455)}=6.03, p<.05$ ) in favor of science teaching students.

### **3.7. Findings Related to the Test-Retest Method**

To determine the reliability of the test-retest method, the PMI was administered to 52 students at 40 days interval, and Pearson product-moment correlation coefficients were calculated. These correlation coefficients were 0.78 for the entire scale, 0.76 for the factor 1, 0.73 for the factor 2, 0.86 for the factor 3, 0.76 for the factor 4, 0.82 for the factor 5 and, 0.85 for the factor 6.

## **4. DISCUSSION and CONCLUSION**

The use of standardized measurement instruments tested in international validity and reliability studies increases the quality of research. Furthermore, the adaptation of scales that are sufficiently known in international publications to Turkish allows researchers obtain comparable data in a shorter time and facilitates communication (Şahin, 1994). From this perspective, PMI developed by Taasobshirazi and Farley (2013), consisting of 24 items and six sub-scales, was adapted to Turkish in the current study.

In order to minimize the differences between cultures in the adaptation process, language and field experts were consulted. CFA was applied to test the structure of the adapted scale. The values obtained from CFA were within the accepted ranges reported in the literature. In other words, the calculated fit indices indicated that the tested model was acceptable. In addition, the moderate- and high-level correlations found between the factors confirmed that divergent validity was achieved. The test-retest scores were calculated to further improve reliability. The t-test conducted between the mean item scores of the upper 27% and lower 27% groups for the discriminative power of the items revealed that the differences were significant for all items. Cronbach's alpha reliability coefficient for the whole scale was calculated as .93.

In conclusion, the factors of the adapted version of the PMI scale and the items under these factors had the same structure as the original PMI. The results obtained from the analyses showed that the adapted PMI had acceptable psychometric values (Klein, 1998). The item distribution was as follows: six items in Factor 1 (knowledge of cognition: declarative, procedural, conditional), four items in Factor 2 (regulation of cognition: information management), four items in Factor 3 (regulation of cognition: monitoring), three items in Factor 4 (regulation of cognition: evaluation), two items in Factor 5 (regulation of cognition: debugging), and five items in Factor 6 (regulation of cognition: planning).

In the Turkish literature, there is no measurement tool that evaluates the physics metacognition of high school and university students. Therefore, it is considered that the adapted PMI will greatly contribute to the field. However, the number of participants and experts was limited to those specified in the method section. In addition, validity of conformity with an equivalent scale was not undertaken. Therefore, the validity of the scale can be further investigated using different scales related to physics and science education. The Turkish PMI presented in the current study can be applied to evaluate the level of high school and university students' metacognition in solving physics problems. It can also be employed to determine the degree to which various methods and techniques affect physics metacognition. In future studies, the validity and reliability analyses of the adapted scale can be retested on data to be obtained from students enrolled in different departments of universities, as well as high school students to increase the generalizability of the adapted scale.

## Acknowledge

This study was supported by the Yozgat Bozok University Research Fund (Project Number: 6602a-EF/18-155). The abstract of this study was presented at 17<sup>th</sup> Classroom Teaching Education Symposium (USOS 2018).

## ORCID

Zeynep Koyunlu Ünlü  <https://orcid.org/0000-0003-3627-1809>

İlbilge Dökme  <https://orcid.org/0000-0003-0227-6193>

## 5. REFERENCES

- Abd-El-Khalick, F., & Akerson, V. (2009). The influence of metacognitive training on preservice elementary teachers' conceptions of nature of science. *International Journal of Science Education*, 31, 2161-2184.
- Akben, N. (2018). Effects of the problem-posing approach on students' problem solving skills and metacognitive awareness in science education. *Research in Science Education*, <https://doi.org/10.1007/s11165-018-9726-7>.
- Anzai, Y., & Yokoyama, T. (1984). Internal models in physics problem solving. *Cognition and Instruction*, 1(4), 397-450.
- Blakey, E., & Spence, S. (1990). *Developing metacognition*. Syracuse, NY: Clearinghouse on Information Resources (ERIC Document Reproduction Service No. ED 327 218). <http://www.nagc.org/index.aspx?id=205> Date of access: 11.01.2018
- Brown, A. L. (1978). Knowing when, where, and how to remember: a problem of metacognition. In R. Glaser (Ed.), *Advances in instructional psychology*, 7, 55-111. New York: Academic Press.
- Büyüköztürk, Ş. (2004). *Sosyal bilimler için veri analizi el kitabı [Handbook of data analysis for social sciences]*. Ankara: Pegem A Yayıncılık.
- Chi, M.T.H. (2006). Two approaches to the study of experts' characteristics. In N. Charness, P.J. Feltovich, R.R. Hoffman, & K.A. Ericsson (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 21-30). New York, NY: Cambridge University Press.
- Colthorpe, K., Sharifirad, T., Ainscough, L., Anderson, S., & Zimbardi, K. (2018). Prompting undergraduate students' metacognition of learning: implementing 'meta-learning' assessment tasks in the biomedical sciences. *Assessment & Evaluation in Higher Education*, 43, 272-285.
- Dianovsky, M. T., & Wink, D. J. (2012). Student learning through journal writing in a general education chemistry course for pre-elementary education majors. *Science Education*, 96, 543-565.
- Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme-I: temel kavramlar ve işlemler [Measurement and scale development in psychology-I: basic concepts and procedures]*. Ankara: Pegem Akademi.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: a new area of cognitive development inquiry. *American Psychologist*, 34(10), 906-911.
- Georghiadis, P. (2004). Making pupils' conceptions of electricity more durable by means of situated metacognition. Research report. *International Journal of Science Education*, 26, 85-99.
- Ghanizadeh, A. (2018). The interplay between reflective thinking, critical thinking, self monitoring, and academic achievement in higher education. *Higher Education*, 74, 101-114.
- Güss, C. D., & Wiley, B. (2007). Metacognition of problem-solving strategies. *Journal of Cognition and Culture*, 7, 1-25.



- Hutner, T. L., & Markman, A. B. (2016). Department-level representations: a new approach to the study of science teacher cognition. *Science Education*, 100(1), 30-56.
- Kurnaz, M. A., & Yiğit, N. (2010). Physics attitude scale: development, validity and reliability. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education*, 4(1), 29-49.
- Lai, E. R. (2011). *Metacognition: a literature review*. Research Reports. <http://www.datec.org.uk/CHAT/chatmetal.htm>. Date of access: 24.12.2017
- Meydan, C. H., & Şeşen, H. (2015). *Yapısal eşitlik modellemesi AMOS uygulamaları [Structural equation modeling AMOS applications]*. Ankara: Detay yayıncılık.
- Neto, A., & Valente, M. O. (1997). *Problem solving in physics: towards a metacognitively developed approach*. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching Oak Brook.
- Nuhoğlu, H., & Yalçın, N. (2004). The development of attitude scale for laboratory and the assessment of preservice teachers' attitudes towards physics laboratory. *Journal of Gazi University Faculty of Education Kirsehir*, 5(2), 317-327.
- Nunnally, J. C. (1978). *Psychometric theory (2nd ed.)*. New York: McGraw-Hill.
- Nunnally, J.C., & Bernstein, I. H. (1994). The assessment of reliability. *Psychometric Theory*, 3, 248-292.
- Öztürk, N. (2017). Assessing metacognition: theory and practices. *International Journal of Assessment Tools in Education*, 4(2), 134-148.
- Patton, M. Q. (2002). *Qualitative research & evaluation methods*. 3rd edition. Sage Publications, Inc.
- Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory Into Practice*, 4, 218-225.
- Rozencaj, P. (2003). Metacognitive factors in scientific problem-solving strategies. *European Journal of Psychology of Education*, 18(3), 281-294.
- Schraw, G., Crippen K. J., & Hartley, K. (2006). Promoting self-regulation in science education: metacognition as part of a broader perspective on learning. *Research in Science Education*, 36, 111-139.
- Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, 19(4), 460-475.
- Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review*, 7(4), 351-371.
- Selcuk, G. S., Caliskan, S., & Erol, M. (2007). The effects of gender and grade levels on Turkish physics teacher candidates' problem solving strategies. *Journal of Turkish Science Education*, 4(1), 92-100.
- Seçer, İ. (2015). *Psikolojik test geliştirme ve uyarlama süreci SPSS ve LISREL uygulamaları [Psychological test development and adaptation process SPSS and LISREL applications]*. Ankara: Anı yayıncılık.
- Şahin, N. (1994). Psikoloji araştırmalarında ölçek kullanımı [Using scale in psychology research]. *Türk Psikoloji Dergisi*, 9(33), 19-26.
- Stewart, J., & Rudolph, J. (2001). Considering the nature of scientific problems when designing science curriculum. *Science Education*, 85, 207-222.
- Taasoobshirazi, G., & Farley, J. (2013). Construct validation of the physics metacognition inventory. *International Journal of Science Education*, 35(3), 447-459.
- Taasoobshirazi, G., Bailey, M., & Farley, J. (2015). Physics metacognition inventory part II: confirmatory factor analysis and rasch analysis. *International Journal of Science Education*, 37(17), 2769-2786.
- Tekbıyık, A., & Akdeniz, A. R. (2010). Ortaöğretim öğrencilerine yönelik güncel fizik tutum ölçeği: geliştirilmesi, geçerlik ve güvenirliği [Physical attitude scale for secondary




- school students: development, validity and reliability]. *The Journal of Turkish Science Education*, 7(4), 134-144.
- Tezbaşaran, A. (1996). *Likert tipi ölçek geliştirme klavuzu [Likert type scale development guide]*. Ankara: Psikologlar Derneği Yayınları.
- Thomas, G. P. (2012). Metacognition in science education: past, present and future considerations. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (vol. 24, pp. 131–144). Dordrecht: Springer.
- Veenman, M. V. J. (2011). Learning to self-monitor and self-regulate. In R. Mayer, & P. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 197-218). New York: Routledge.
- Veenman, M. V. J., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: age and task differences. *Learning & Individual Differences*, 15(2), 159-176.
- Zohara A., & Barzilai, S. (2013). A review of research on metacognition in science education: current and future directions. *Studies in Science Education*, 49, 121-169.

**Appendix 1. The Turkish Version of PMI**

| <b>Maddeler</b>  | <b>Tamamen katılıyorum</b> | <b>Katılıyorum</b> | <b>Kararsızım</b> | <b>Katılmıyorum</b> | <b>Hiç katılmıyorum</b> |
|--|----------------------------|--------------------|-------------------|---------------------|-------------------------|
| 1. Bir fizik problemini çözmeye başlamadan önce problemin ne istediği hakkında düşünürüm.                                  |                            |                    |                   |                     |                         |
| 2. Bir fizik problemini çözerken hedeflerime ulaşip ulaşmadığımı belirli aralıklarla kendi kendime sorarım.                |                            |                    |                   |                     |                         |
| 3. Bir fizik problemini anlamadığımda yardım isterim.  |                            |                    |                   |                     |                         |
| 4. Fizik problemlerini çözmemde yardımcı olması için serbest cisim diyagramları çizerim.                                   |                            |                    |                   |                     |                         |
| 5. Fizik problemlerini ne kadar iyi çözebildiğim konusunda sağlıklı bir değerlendirme yaparım.                             |                            |                    |                   |                     |                         |
| 6. Fizik problemlerini çözerken, elimden gelenin en iyisini nasıl yapacağımı bilirim.                                      |                            |                    |                   |                     |                         |
| 7. Fizik problemlerini çözerken, kullandığım her bir stratejiye özgü belirli bir amacım vardır.                            |                            |                    |                   |                     |                         |
| 8. Bir fizik problemini çözdükten sonra geriye dönüp çözümümü kontrol ederim.  |                            |                    |                   |                     |                         |
| 9. Bir fizik problemini çözdükten sonra, cevabımı kontrol ederim.  |                            |                    |                   |                     |                         |
| 10. Fizik problemlerini çözmeye bana yardımcı olacak serbest cisim diyagramları kullanırım.                                |                            |                    |                   |                     |                         |
| 11. Bir fizik problemini çözerken, problemi doğru çözmek için gereken stratejiyi nasıl kullanacağımı bilirim.              |                            |                    |                   |                     |                         |
| 12. Bir fizik problemini çözerken, belirli bir stratejiyi hangi sebeple kullanacağımı bilirim.                             |                            |                    |                   |                     |                         |
| 13. Bir fizik problemini çözerken, belli bir stratejiyi ne zaman kullanacağımı bilirim.                                    |                            |                    |                   |                     |                         |
| 14. Bir fizik problemini çözmeden önce, sonucun ne çıkabileceğini yaklaşık olarak tahmin ederim.                           |                            |                    |                   |                     |                         |
| 15. Bir fizik problemini çözerken, problemi ne kadar doğru çözüyor olduğuma dair kendi kendime sorular sorarım.            |                            |                    |                   |                     |                         |
| 16. Bir fizik problemini çözerken, problemi ne kadar doğru çözüyor olduğumu belirli aralıklarla değerlendiririm.           |                            |                    |                   |                     |                         |
| 17. Bir fizik problemini çözdükten sonra, doğru yöntemleri uygulayıp uygulamadığımı görmek için çözümümü gözden geçiririm. |                            |                    |                   |                     |                         |
| 18. Fizik problemlerinin çözümü için serbest cisim diyagramlarının neden önemli olduğunu bilirim.                          |                            |                    |                   |                     |                         |
| 19. Bir fizik problemini çözmeye başlamadan önce, problemi nasıl çözeceğimi planlarım.                                     |                            |                    |                   |                     |                         |
| 20. Bir fizik problemini çözmeden önce problemin önemli kısımlarının tamamını tespit ederim.                               |                            |                    |                   |                     |                         |
| 21. Bir fizik problemini çözerken, hedeflerime ulaşip ulaşmadığımı kendi kendime sorarım.                                  |                            |                    |                   |                     |                         |
| 22. Çözmeye çalıştığım fizik problemlerini anlamadığım zaman yardım isterim.   |                            |                    |                   |                     |                         |
| 23. Bir fizik problemini çözerken serbest cisim diyagramları çizerim.  |                            |                    |                   |                     |                         |
| 24. Bir fizik problemini çözmeden önce problemde gerek duymadığım bilgileri elim.  |                            |                    |                   |                     |                         |

## Performance Evaluation Using the Discrete Choquet Integral: Higher Education Sector

Seher Nur Sülkü<sup>1</sup>, Deniz Koçak <sup>1,\*</sup>

<sup>1</sup> Department of Econometrics, Ankara Hacı Bayram Veli University, Ankara, Turkey

### ARTICLE HISTORY

Received: 14 November 2018

Accepted: 05 March 2019

### KEYWORDS

Performance evaluation,  
Fuzzy measure,  
Discrete Choquet integral,  
*k*-means

**Abstract:** Performance evaluation functions as an essential tool for decision makers in the field of measuring and assessing the performance under the multiple evaluation criteria aspect of the systems such as management, economy, and education system. Besides, academic performance evaluation is one of the critical issues in higher institution of learning. Even though the academic evaluation criteria are inherently dependent, most of the traditional evaluation methods take no account of the dependency. Currently, the discrete Choquet integral can be proposed as a useful and effective aggregation operator due to being capable of considering the interactions among the evaluation criteria. In this paper, it is aimed to solve an academic performance evaluation problem of students in a university in Turkey using the discrete Choquet integral with the complexity-based method and the entropy-based method. Moreover, the *k*-means method, which has been widely used for evaluating students' performance over 50 years, is used to compare the effectiveness and the success of two different frameworks based on discrete Choquet integral in the robustness check. Our results indicate that the entropy-based Choquet integral outperforms the complexity-based Choquet and *k*-means method in most of the cases.

## 1. INTRODUCTION

In recent years, performance evaluation plays an important role due to the lack of operational tools provided objective information in the managerial, educational, and economic areas. Therefore, performance evaluation can be seen as a tool developed for determining whether the wide-ranging set of evaluation criteria is met in the associated areas. Conversely, academic performance evaluation is one of the critical issues in higher institution of learning. Based on this critical issue, many traditional evaluation techniques, which are mainly based on the weighted arithmetic mean, have been widely used, but these techniques only consider situations where all the evaluation criteria are independent. Contrary to the weighted arithmetic mean, the Choquet integral is an appropriate substitute that allows to capture dependency among evaluation criteria (Marichal & Roubens, 2000). The Choquet integral introduced by Choquet

---

CONTACT: Deniz Koçak ✉ [denizkocak36@gmail.com](mailto:denizkocak36@gmail.com) 📍 Department of Econometrics, Ankara Hacı Bayram Veli University, Ankara, Turkey

ISSN-e: 2148-7456 / © IJATE 2019

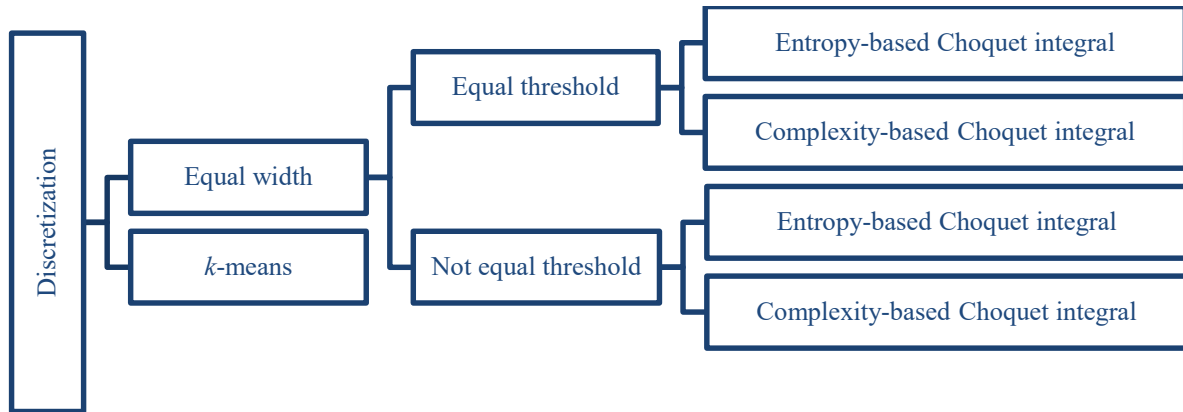
is an aggregation operator that is extensively employed in quantitative problems such as multi-criteria and multi-objective optimization problems, economics problems, and multi-regression problems, etc. (Choquet 1954; Cui & Li 2008; Angilella et al., 2017). Moreover, the Choquet integral provides an indirect method that reflects the relative importance of evaluation criteria, dependency among them, and their ordered positions in these problems (Angilella et al., 2015; Xu 2010).

Early 2000s, the data mining techniques have been used in the educational area and Educational Data Mining (EDM) has emerged (Baker & Yacef 2009; Peña-Ayala, 2014). In recent years, the tools of the EDM are widely used with educational data (Slater et al., 2017). The new operational tools that serve accountability policies have emerged (Huber & Skedsmo 2016). However, the research in educational data mining have generated the need for rethinking of these new operational tools in handling dependent evaluation criteria. Besides, it is established that more research is needed to specify educational goals for a valid evaluation of students' skills (Herde et al., 2016). In recent years, Shieh, Wu and Liu (2009) proposed discrete Choquet integral with a complexity-based method to evaluate students' performance where the discrete Choquet integral is an adequate aggregation operator which takes the interactions into account. Chang, Liu, Tseng and Chang (2009) found out the poor performance of the traditional regression models in the evaluation of the students' performance when there are interactions among the attributes with using a real data set from a junior high school; and then showed that multiple-mutual information based Choquet integral regression models provide better performance while comparing the joint entropy based and complexity based Choquet integral. In another study, Wang, Nian, Chu and Shi (2012) used the nonlinear multi-regression based on the Choquet integral in order to evaluate the final grade of the students considering previous records such as scores of tests, the average score of quizzes, the number of absent class meeting and the number of incomplete homework as interactive predictive attributes. Branke, Correnre, Greco, Slowinski and Zielniewicz (2016) used Choquet integral as a preference model and suggested an interactive multiobjective evolutionary algorithm.

The discrete Choquet integral has been newly started to be preferred by the researchers due to their success in terms of considering the evaluation criteria dependency. The method is an important kind of non-additive integrals (Wang & Ha 2008), and nowadays its theory is applied by the authors in decision making problems (Grabisch, 1996). Nevertheless, we encountered that there is still a limited number of studies in this context. Only the mentioned studies take the interaction among criteria into account in the literature of academic performance evaluation. Therefore, the purpose of this study is to use various discretization methods and the discrete Choquet integral in order to provide realistic evaluation in educational system. More precisely, the academic performance of students from a university in Turkey are evaluated employing both the entropy-based and the complexity-based discrete Choquet integral and the  $k$ -means method. Thereafter, the effectiveness and success of the different discretization techniques are compared, and the model evaluation of these different methods is carried out. The steps of the present analysis are summarized in [Figure 1](#).

In discretization process, a nonoverlapping partition of a continuous domain is obtained. For this aim, first of all continuous attributes are sorted and then the number of intervals are defined. For example, if there will be  $k$  intervals then there will be  $k-1$  split points. Thus, a researcher actually defines intervals by deciding on the place of split points. Thereafter, all continuous attributes falling into the same interval are automatically mapped to the same categorical value. Hence, the key task is finding meaningful intervals in discretization (Kononenko and Kukar 2007). The equal width interval methods divide the continuous data into the categorical data by using user specified number of intervals. In case of "equal threshold" of the equal width interval methods, if there are  $n \times 1$  vectors consisting of three continuous variables, i.e.  $X$ ,  $Y$  and  $Z$ , the

data matrix is obtained by assigning the same threshold value to all of them. On the contrary, in case of “not equal threshold”, the data matrix was obtained by assigning a different threshold for  $X$ ,  $Y$  and  $Z$ . Then the entropy and complexity based methods are applied to this matrix. The results of these methods are intermingled with the discrete Choquet integral.



**Figure 1.** Overview of the discretization methods

Besides, the  $k$ -means method is used to compare the effectiveness and the success of two different frameworks based on discrete Choquet integral in the robustness check. Regardless of the fact that the method was presented many years ago, it is one of the most widespread classification algorithms and widely used for evaluating students' performance in educational data mining (Veeramuthu et al., 2014; Jain, 2010). For this reason, the  $k$ -means method is not explained technically, but its results in the robustness check is presented.

In this study, the aim is to provide a sufficient and comprehensible background on the discrete Choquet integral method, thus the empirical analysis of the study is exemplified step by step. It is believed that a reader who is even unfamiliar to the Choquet integral methodology can redo the present analyses following the steps which are explained thoroughly in the main text. The rest of this paper is organized as follows. Section 2, a brief introduction of the the discretization techniques, outline of the the fuzzy measure, and the discrete Choquet integrals with entropy-based and complexity-based constructs are presented. The research findings and the robustness check results are presented and discussed in Section 3. Finally, Section 4 concludes the study.

## 2. METHOD

### 2.1. Discretization

The evaluation of the academic performance can be considered as a multi-criteria decision making (MCDM) problem. In these problem refers to the evolution of a partition matrix of a data set, and describing the component of a data set from the most preferred alternatives to the least preferred alternatives (Zopounidis & Doumpos, 2002). In many real-life decision making problems that have multi criteria, it is important to preprocess data to effectively apply the algorithms (Kononenko et al., 2007).

Preprocessing the data has a number of steps such as data transformation, cleaning, and data reduction (Pyle, 1999). Currently, discretization is one of the most popular reduction techniques (Garcia et al., 2013). The aim of discretization is to transform continuous attributes which take infinitely many values into categorical attributes and which are significantly reduced subset of discrete values to make the representation of information easier and to learn from the data more accurately and fast (Liu et al., 2002). The discretization methods are summarized in Table 1 (Dougherty et al.,1995).

Detailed review on the discretization methods can be found in Garcia et al., (2013) and Liu et al., (2002). The main separation between discretization methods is whether the class information is employed or not. In the supervised discretization, the class information is considered in the classification but not in unsupervised discretization. Another distinction between discretization methods is global versus local discretization. Global discretization methods use the complete instance space to discretize whereas local discretization methods use only a region of the instance space (Chmielewski & Grzymala-Busse 1996).

The basic unsupervised methods, equal frequency and equal width, do not perform well when there are outliers in the data and when continuous attributes do not follow the uniform distribution (Tan et al., 2005; Catlett, 1991). To deal with these shortcomings, supervised discretization methods have been developed and class information is used to establish the appropriate intervals. There are not as many unsupervised methods as supervised methods, that may be related to the fact that discretization is usually related with the classification task. However, if the class information is not available, only unsupervised methods can be used.

**Table 1.** Summary of discretization methods

|              | Global  | Local   |
|--------------|---|---|
| Supervised   | 1RD<br>Adaptive Quantizers<br>Chi Merge (Kerber)<br>D-2 (Catlett)<br>Fayyad and Irani / Ting<br>Supervised MCC<br>Predictive Value Max. | Vector Quantization<br>Hierarchical Maximum Entropy<br>Fayyad and Irani<br>C4.5 |
| Unsupervised | Equal width interval<br>Equal frequency interval<br>Unsupervised MCC  | <i>k</i> -means clustering  |

The unsupervised discretization methods can be regarded as sorting problems or separating problems that distinguish the probability occurrences from a mixing of probability laws (Potzelberger & Felsenstein 1993). However, in these methods, the aggregation operators are needed for the fusion of several input values into a single output value (Calvo et al., 2002). In this respect, the discrete Choquet integral is a suitable aggregation operator by taking into the dependency among criteria account (Wen et al., 2016). Besides, the Choquet integral is remarkable in terms of modeling specific interactions of such a broad spectrum of topics including education, health, living conditions (Kasparian & Rolland 2012).

## 2.2. Fuzzy measure and the discrete Choquet integral

The definitions of fuzzy measures and Choquet integral are as follows (Shieh et al., 2009):

*Definition 1.* Let  $N$  be a finite set of criteria and  $P(N)$  be the power set of  $N$ . A discrete fuzzy measure ( $\mu$ ) on  $N$  is a set function  $\mu: 2^N \rightarrow [0,1]$  which satisfies the following axioms. Besides,  $\forall S \subseteq N, \mu(S)$  can be explained as the weight of the coalition  $S$ .

- (1)  $\mu(\emptyset) = 0, \mu(N) = 1$  (boundary condition)
- (2)  $A \subseteq B \Rightarrow \mu(A) \leq \mu(B), A, B \in P(N)$  (monotonicity)

*Definition 2.* Let  $\mu$  be a fuzzy measure on  $N = \{1, 2, \dots, n\}$ . The discrete Choquet integral of  $x$  in connection with  $\mu$  is defined as:



$$C_v = \sum_{i=1}^n x_{(i)} [\mu(A_{(i)}) - \mu(A_{(i+1)})], \quad (1)$$

where  $(.)$  implies a permutation on  $N$  such that  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . Additionally,  $A_{(i)} = \{(i), (i+1), \dots, (n)\}$  and  $A_{(n+1)} = \phi$ .

There is a need for fuzzy measure to calculate the discrete Choquet integral. In this paper, the complexity based and the entropy based fuzzy measure are qualified to be fuzzy measures. The detailed definition of the measures which needs to gratify the fuzzy measure axioms, is given below:

*Definition 3.* The complexity  $C$  of a discrete random variable  $N$  is defined as the function which counts the number of different forms in  $N$ .  $C_1$ , is defined as equation (2).  $\forall S \subseteq N$ , to calculate the complexity of the subsets of criteria of  $N$ . Clearly,  $C_1(\phi) = 0$  and if  $A \subseteq B \Rightarrow C_1(A) \leq C_1(B)$ ,  $A, B \in N$ . That is  $C_1$ , is a fuzzy measure.

$$C_1(S) = \frac{C(S)}{C(N)}, \quad (2)$$

*Definition 4.* Let  $A$  be a discrete random variable and  $p^A$  be the probability of  $A$ , then the entropy of  $A$  is defined as:

$$h(A) = - \sum p^A \log_2 p^A, p^A > 0. \quad (3)$$

Let  $B$  be a discrete random vector which contains at least two discrete random variables,  $p^B$  be the joint probability and  $h(B)$  the joint entropy. By using the idea of the joint entropy to calculate the entropy of the subsets of criteria of  $N$ , the fuzzy measure  $(\mu_1)$  is defined as:

$$\mu_1(S) = \frac{h(S)}{h(N)}, \forall S \subseteq N. \quad (4)$$

### 2.3. Evaluation the performance of the models

Usually practical applications that used the entropy-based and the complexity-based discrete Choquet integral evaluate the performance of the models with a metric called as ‘‘accuracy’’. Furthermore, in the applications of  $k$ -means method, the cluster evaluations can be done with the measures of cluster cohesion and cluster separation (Tan et al., 2005). However, when different discretization techniques and their different model evaluation methods are compared, the mean square error ( $MSE$ ) criteria would be more suitable to choose the best performing one among them (Greene, 2016). In this study,  $MSE$  was employed to evaluate alternative models performances. While comparing the models, as  $MSE$  gets smaller, the model does better performance. Thus, the model with the smallest  $MSE$  value is preferred. Let  $\theta$  be a parameter and  $\hat{\theta}$  an estimator of this parameter, the mean square error of an estimator is defined as below:

$$MSE [\hat{\theta}|\theta] = E [(\theta - \hat{\theta})^2]. \quad (5)$$

### 3. EMPIRICAL STUDY and RESULTS

The raw data set shown in Table 2 is composed of 33 students' course scores from Econometrics Department at Gazi University. The courses are chosen as follows: Introduction to Statistics and Probability-II ( $D_1$ ), Microeconomics ( $D_2$ ), Macroeconomics ( $D_3$ ), Mathematics-II ( $D_4$ ), and Econometrics-I ( $EKON$ ).

The  $EKON$  scores of the students are set as control group in the analysis because Econometrics-I is a discipline that requires comprehensive knowledge of the other four courses. Besides, the minimum and maximum score for each course are 1 and 100, respectively.

In the empirical study of this paper, it is aimed to estimate the Econometrics scores of the students with using the students' scores of Introduction to Statistics and Probability-II, Microeconomics, Macroeconomics and Mathematics-II courses. For this aim, the discrete Choquet integral was used as an aggregation and estimation operator because of the fact that there are interactions among these four courses. Thereafter, to measure the success of the estimation based on the Choquet integral, the mean square error was computed by using the students' raw scores of Econometrics-I (see Table 2), and the estimation scores (see Table 7).

**Table 2.** Raw data scores of the students

| Student | $D_1$ | $D_2$ | $D_3$ | $D_4$ | $EKON$ | Student | $D_1$ | $D_2$ | $D_3$ | $D_4$ | $EKON$ |
|---------|-------|-------|-------|-------|--------|---------|-------|-------|-------|-------|--------|
| 1       | 55.8  | 42    | 52    | 76    | 30     | 18      | 62.6  | 66.2  | 66    | 90    | 65.2   |
| 2       | 42    | 63.8  | 49    | 94    | 38     | 19      | 66    | 28.8  | 45    | 78    | 49.4   |
| 3       | 39.6  | 45    | 52    | 50.2  | 68     | 20      | 68    | 45    | 73    | 100   | 38.8   |
| 4       | 40.4  | 42    | 51    | 94    | 61.4   | 21      | 60    | 47.2  | 51    | 100   | 48     |
| 5       | 61.6  | 54.6  | 56    | 86    | 77.8   | 22      | 66    | 41.8  | 47    | 92    | 44     |
| 6       | 67.8  | 45    | 77    | 100   | 57     | 23      | 79.2  | 58.2  | 71    | 100   | 66.2   |
| 7       | 36.8  | 47.2  | 46    | 90    | 36.8   | 24      | 29.4  | 41.8  | 61    | 73    | 37.6   |
| 8       | 52    | 57    | 54    | 87    | 37.2   | 25      | 68.4  | 50.4  | 79    | 92    | 51.6   |
| 9       | 44.8  | 45    | 59    | 100   | 59.8   | 26      | 53.8  | 34.8  | 59    | 93    | 51.8   |
| 10      | 32.6  | 44.8  | 45    | 86    | 32.6   | 27      | 74    | 47.8  | 74    | 100   | 73.2   |
| 11      | 62.2  | 48    | 50.4  | 94    | 43.2   | 28      | 46.2  | 49.8  | 47    | 74    | 19.8   |
| 12      | 67.4  | 57.8  | 55    | 100   | 39.2   | 29      | 68.2  | 47.8  | 59    | 100   | 60.6   |
| 13      | 64    | 52.2  | 45    | 49    | 67.8   | 30      | 76.8  | 72    | 94    | 100   | 83.6   |
| 14      | 54    | 13.2  | 62    | 78    | 2.8    | 31      | 56    | 31.6  | 47    | 90    | 33.6   |
| 15      | 50.4  | 22.4  | 47    | 74    | 41.8   | 32      | 76.8  | 55.6  | 78    | 96    | 72     |
| 16      | 67.6  | 53.6  | 57    | 94    | 50     | 33      | 72.8  | 20.2  | 53    | 56.6  | 75.2   |
| 17      | 63.4  | 42    | 45    | 97    | 51     |         |       |       |       |       |        |

First of all, the descriptive statistics and the normality of the data were checked out. As presented in the Table 3, the average of  $EKON$  is 50.46 while the averages of  $D_1$  and  $D_3$  are around 60, the average of  $D_2$  is almost 46. The mathematics course has the highest average, almost 88. Since  $n = 33$ , Kolmogorov-Smirnov test and Jarque Bera test are appropriate for testing normality. With respect to the Jarque Bera test, the null hypothesis of normality for the distribution of returns is rejected at the significance level of 5% and all variables are not normally distributed. Furthermore, according to Kolmogorov-Smirnov test,  $D_2$ ,  $D_3$  and  $D_4$  variables are not normally distributed;  $D_1$  and  $EKON$  variables are normally distributed (Asymptotic Significance > 0.05).

**Table 3.** Results of one-sample Kolmogorov-Smirnov test and Jarque Bera

|                                    |                | $D_1$ | $D_2$  | $D_3$ | $D_4$ | $EKON$ |
|------------------------------------|----------------|-------|--------|-------|-------|--------|
| Normal Parameters                  | Mean           | 58.38 | 45.90  | 57.77 | 87.39 | 50.46  |
|                                    | Std. Deviation | 13.38 | 12.67  | 12.38 | 14.30 | 17.87  |
| Most Extreme Differences           | Positive       | 0.08  | 0.08   | 0.16  | 0.19  | 0.08   |
|                                    | Negative       | -0.14 | -0.191 | -0.15 | -0.21 | -0.07  |
| Test Statistic                     |                | 0.14  | 0.191  | 0.16  | 0.21  | 0.08   |
| Asymptotic Significance (2-tailed) |                | 0.09  | 0.00   | 0.04  | 0.00  | 0.20   |
| Skewness                           |                | -0.54 | -0.58  | 1.16  | -1.44 | -0.32  |
| Kurtosis                           |                | -0.59 | 0.86   | 0.84  | 1.536 | 0.21   |
| Jarque Bera                        |                | 19.32 | 8.13   | 13.77 | 14.40 | 11.26  |

Before applying the complexity-based and entropy-based methods, the number of the level of score ( $m$ ) which transforms the continuous raw data into the categorical level of the score should be decided on. This level of score can be defined by the users or can be stated in terms of the interval width for the equal width interval method.

**Table 4.** Categorical data scores of the students ( $m = 3$ )

| Student | $D_1$ | $D_2$ | $D_3$ | $D_4$ | $EKON$ | Student | $D_1$ | $D_2$ | $D_3$ | $D_4$ | $EKON$ |
|---------|-------|-------|-------|-------|--------|---------|-------|-------|-------|-------|--------|
| 1       | 2     | 2     | 1     | 2     | 2      | 18      | 3     | 3     | 2     | 3     | 3      |
| 2       | 1     | 3     | 1     | 3     | 2      | 19      | 3     | 1     | 1     | 2     | 2      |
| 3       | 1     | 2     | 1     | 1     | 3      | 20      | 3     | 2     | 2     | 3     | 2      |
| 4       | 1     | 2     | 1     | 3     | 3      | 21      | 2     | 2     | 1     | 3     | 2      |
| 5       | 2     | 3     | 1     | 3     | 3      | 22      | 3     | 2     | 1     | 3     | 2      |
| 6       | 3     | 2     | 2     | 3     | 3      | 23      | 3     | 3     | 2     | 3     | 3      |
| 7       | 1     | 2     | 1     | 3     | 2      | 24      | 1     | 2     | 1     | 2     | 2      |
| 8       | 2     | 3     | 1     | 3     | 2      | 25      | 3     | 2     | 3     | 3     | 2      |
| 9       | 1     | 2     | 1     | 3     | 3      | 26      | 2     | 2     | 1     | 3     | 2      |
| 10      | 1     | 2     | 1     | 3     | 2      | 27      | 3     | 2     | 2     | 3     | 3      |
| 11      | 2     | 2     | 1     | 3     | 2      | 28      | 2     | 2     | 1     | 2     | 1      |
| 12      | 3     | 3     | 1     | 3     | 2      | 29      | 3     | 2     | 1     | 3     | 3      |
| 13      | 3     | 2     | 1     | 1     | 3      | 30      | 3     | 3     | 3     | 3     | 3      |
| 14      | 2     | 1     | 2     | 2     | 1      | 31      | 2     | 1     | 1     | 3     | 2      |
| 15      | 2     | 1     | 1     | 2     | 2      | 32      | 3     | 3     | 3     | 3     | 3      |
| 16      | 3     | 3     | 1     | 3     | 2      | 33      | 3     | 1     | 1     | 1     | 3      |
| 17      | 3     | 2     | 1     | 3     | 2      |         |       |       |       |       |        |

First of all, equal thresholds approach of the equal width interval method was used. The equal width interval method converts the continuous data into the categorical data by employing user specified number of intervals. Here the number of intervals as  $m = 2, 3, 4, 5, 6, 7, 8,$  and  $9$  were specified. Thereafter, the raw data in Table 2 was transformed by using “hist.m” program of Matlab for  $D_1, D_2, D_3, D_4$  and  $EKON$  variables when  $m = 2, 3, 4, 5, 6, 7, 8,$  and  $9$ . Later, the complexity-based and entropy-based fuzzy measure were computed at each level of score ( $m = 2, 3, 4, 5, 6, 7, 8,$  and  $9$ ) with applying the equations (1), (2), and (3) to determine the dependency of the evaluation criteria. Final identified fuzzy measures for each subset were computed by Matlab and showed in Table 5. Before presenting the Table 5, in order to make clear that how the final values are obtained  $m=3$  case was provided as an example. Here, how each of the steps was followed when  $m=3$  was employed is summarized in the preceding

paragraph. Firstly, continuous raw data scores (in Table 2) were converted into categorical data. When  $m=3$  is employed, the categorical data score for each course for each student can be 1, 2 or 3. Table 4 shows the categorical data scores for each criterion transformed from the raw data scores by using “hist.m” program of Matlab.

Furthermore, the histograms of the  $D_1, D_2, D_3, D_4$  and  $EKON$  courses when the number of the level score is equal to three,  $m=3$ , can be seen in Figure 2. For example, for Microeconomics ( $D_2$ ) course, students with grade in the interval of  $[0, 32.8)$  constitute the first category and each observation in this group takes categorical value “1”, students with grade in the interval of  $[32.8, 52.4)$  constitute the second category, and each observation in this group takes categorical value “2” and students with grade in the interval of  $[52.4, 72)$  constitute the third category and each observation in this group takes categorical value “3”.

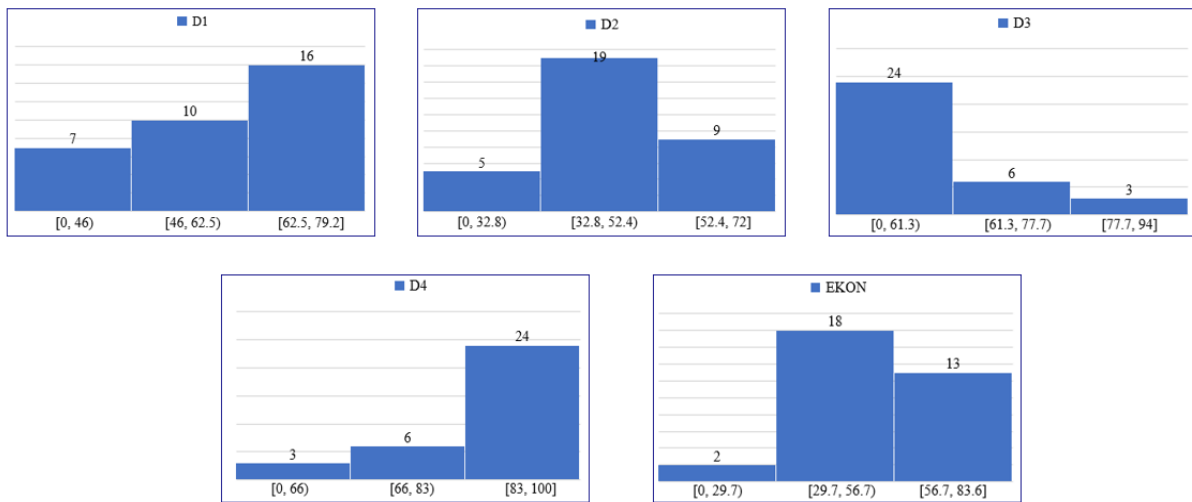


Figure 2. Histograms of the courses for  $m = 3$

For instance, in Table 2, the first student’s grade for  $D_2$  is 42, so this student belongs to second category and in Table 4 in the column of  $D_2$  this observation takes value “2”. For each course raw data of grades are converted into categorical data in the same manner. For each histogram of the courses, the first column shows how many times “1” value is repeated, second column shows how many times “2” value is repeated, and the third column shows how many times “3” value is repeated. Besides, the numbers at which intervals correspond to these values are shown below the columns. Now, to obtain entropy based fuzzy measure,  $h(N)$  was computed. When the transformed data scores of the students are considered, there are 19 different joint pattern in Table 4 these are:  $(2,2,1,2), (1,3,1,3), (1,2,1,1), (1,2,1,3), (2,3,1,3), (3,2,2,3), (2,2,1,3), (3,3,1,3), (3,2,1,1), (2,1,2,2), (2,1,1,2), (3,2,1,3), (3,3,2,3), (3,1,1,2), (1,2,1,2), (3,2,3,3), (3,3,3,3), (2,1,1,3),$  and  $(3,1,1,1)$ . Besides, how many times the patterns are repeated are given respectively 2, 1, 1, 4, 2, 3, 2, 2, 1, 1, 1, 3, 3, 1, 1, 1, 2, 1, and 1. It means that in Table 4  $(2,2,1,2)$  is repeated twice,  $(1,3,1,3)$  is repeated once, and so on. Thus, the joint probabilities are defined and then the entropy of the finite set of criteria ( $N$ ) employing equation 3 could be calculated as:

$$\begin{aligned}
 h(N) &= - \sum p \log_2 p \\
 &= -0.06 * \log_2(0.06) - 0.03 * \log_2(0.03) - 0.03 * \log_2(0.03) - 0.12 * \log_2(0.12) - 0.06 * \log_2(0.06) - 0.09 * \\
 &\quad \log_2(0.09) - 0.06 * \log_2(0.06) - 0.06 * \log_2(0.06) - 0.03 * \log_2(0.03) - 0.03 * \log_2(0.03) - 0.03 * \\
 &\quad \log_2(0.03) - 0.03 * \log_2(0.03) - 0.09 * \log_2(0.09) - 0.09 * \log_2(0.09) - 0.03 * \log_2(0.03) - 0.03 * \\
 &\quad \log_2(0.03) - 0.03 * \log_2(0.03) - 0.06 * \log_2(0.06) - 0.03 * \log_2(0.03) - 0.03 * \log_2(0.03) \\
 &= 4.07
 \end{aligned}$$

Now the subsets of criteria of  $N$  which are  $\forall S \subseteq N$  was introduced: empty set,  $\{D_1\}$ ,  $\{D_2\}$ ,  $\{D_3\}$ ,  $\{D_4\}$ ,  $\{D_1, D_2\}$ ,  $\{D_1, D_3\}$ ,  $\{D_1, D_4\}$ ,  $\{D_2, D_3\}$ ,  $\{D_2, D_4\}$ ,  $\{D_3, D_4\}$ ,  $\{D_1, D_2, D_3\}$ ,  $\{D_1, D_2, D_4\}$ ,  $\{D_1, D_3, D_4\}$ ,  $\{D_2, D_3, D_4\}$ , and  $\{D_1, D_2, D_3, D_4\}$ . These subsets were symbolized as respectively: (0,0,0,0), (1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1), (1,1,0,0), (1,0,1,0), (1,0,0,1), (0,1,1,0), (0,1,0,1), (0,0,1,1), (1,1,1,0), (1,1,0,1), (1,0,1,1), (0,1,1,1), and (1,1,1,1) as shown in Table 5. For example, the effect of the only  $\{D_1\}$  course is known, that situation is symbolized as (1,0,0,0); when the effect of the  $\{D_1, D_2\}$  courses is known, that situation is symbolized as (1,1,0,0). Then the entropy of the subsets of criteria of  $N$ , i.e.  $h(S)$  is calculated using equation 3. For example in order to calculate  $h(D_1)$  Table 4 is considered and the column of  $D_1$  is observed to see how many times “1”, “2” and “3” categories are repeated; “1” is repeated 7 times, “2” is repeated 10 times, and “3” is repeated 16 times.  $h(D_1)$  is calculated as follow:

$$h(D_1) = -\frac{7}{33} * \log_2 \left(\frac{7}{33}\right) - \frac{10}{33} * \log_2 \left(\frac{10}{33}\right) - \frac{16}{33} * \log_2 \left(\frac{16}{33}\right) = 1.50$$

For instance, if  $h(D_1, D_2)$  is considered,  $D_1$  and  $D_2$  columns are simultaneously examined and it is seen that “2, 2” case appears five times, “1, 3” once, “1, 2” six times and so on, thus:

$$\begin{aligned} h(D_1, D_2) &= -\frac{5}{33} * \log_2 \left(\frac{5}{33}\right) - \frac{1}{33} * \log_2 \left(\frac{1}{33}\right) - \frac{6}{33} * \log_2 \left(\frac{6}{33}\right) - \frac{2}{33} * \log_2 \left(\frac{2}{33}\right) - \frac{8}{33} \\ &\quad * \log_2 \left(\frac{8}{33}\right) - \frac{7}{33} * \log_2 \left(\frac{7}{33}\right) - \frac{3}{33} * \log_2 \left(\frac{3}{33}\right) - \frac{2}{33} * \log_2 \left(\frac{2}{33}\right) \\ &= 2.76 \end{aligned}$$

Thus, the entropies of the selected subsets as an example are calculated as follow:

$$\begin{aligned} h(D_1) &= 1.50 \\ h(D_1, D_2) &= 2.76 \\ h(D_1, D_2, D_3) &= 3.50 \\ h(D_1, D_2, D_3, D_4) &= 4.07 \end{aligned}$$

Now, the fuzzy measures can be obtained by employing equation 4 as  $\mu_1(S) = \frac{h(S)}{h(N)}$ , ( $\forall S \subseteq N$ ). As shown in Table 5 for  $m=3$ , the entropy based fuzzy measures for the selected subsets as an example are defined as follows. Besides, the entropy based fuzzy measure of the empty set is always equal to 0.

$$\begin{aligned} \mu_1(D_1) &= \frac{h(D_1)}{h(N)} = \frac{1.50}{4.07} = 0.37 \\ \mu_1(D_1, D_2) &= \frac{h(D_1, D_2)}{h(N)} = \frac{2.76}{4.07} = 0.68 \\ \mu_1(D_1, D_2, D_3) &= \frac{h(D_1, D_2, D_3)}{h(N)} = \frac{3.50}{4.07} = 0.86 \\ \mu_1(D_1, D_2, D_3, D_4) &= \frac{h(D_1, D_2, D_3, D_4)}{h(N)} = \frac{4.07}{4.07} = 1 \end{aligned}$$

The entropy based fuzzy measures for  $m=3$  is obtained, and then the complexity based fuzz measures is obtained. Firstly, the complexity of the discrete random variable, i.e.  $C(N)$  is needed to be computed in equation 2. When the transformed data scores of the students were considered, there was 19 different joint pattern i.e., (2,2,1,2), (1,3,1,3), (1,2,1,1), (1,2,1,3), (2,3,1,3), (3,2,2,3), (2,2,1,3), (3,3,1,3), (3,2,1,1), (2,1,2,2), (2,1,1,2), (3,2,1,3), (3,3,2,3), (3,1,1,2), (1,2,1,2), (3,2,3,3), (3,3,3,3), (2113), and (3,1,1,1) (see Table 4). Thus, through the complexity counts the number of different pattern is  $C(N) = 19$ .

Thereafter, the complexity of the subsets of criteria of  $N$ , i.e.  $C(S)$  is calculated. For instance, there are three features in  $D_1$ : 1,2,3; there are three features in  $D_2$ : 1,2,3; there are three features in  $D_3$ : 1,2,3; thus, the complexities of the selected subsets as an example are calculated as follow:

$$\begin{aligned} C(D_1) &= 3 \\ C(D_1, D_2) &= 8 \\ C(D_1, D_2, D_3) &= 13 \\ C(D_1, D_2, D_3, D_4) &= 19 \end{aligned}$$

Similarly, after the complexity for each subset of  $N$  is calculated, the complexity based fuzzy measures can be obtained by employing equation 2 as  $C_1(S) = \frac{C(S)}{C(N)}$ , ( $\forall S \subseteq N$ ). The complexity based fuzzy measures for the selected subsets as an example are computed for  $m=3$  as follows and the results are given in Table 5. Besides, the complexity based fuzzy measure of the empty set is always equal to 0.

$$\begin{aligned} C_1(D_1) &= \frac{C(D_1)}{C(N)} = \frac{3}{19} = 0.16 \\ C_1(D_1, D_2) &= \frac{C(D_1, D_2)}{C(N)} = \frac{8}{19} = 0.42 \\ C_1(D_1, D_2, D_3) &= \frac{C(D_1, D_2, D_3)}{C(N)} = \frac{13}{19} = 0.68 \\ C_1(D_1, D_2, D_3, D_4) &= \frac{C(D_1, D_2, D_3, D_4)}{C(N)} = \frac{19}{19} = 1 \end{aligned}$$

Up to now, how the entropy and complexity based fuzzy measures are achieved for  $m=3$  have been explained. These values are computed for each level of score ( $m = 2, 3, 4, 5, 6, 7, 8$ , and 9) in the same manner. Finally, the identified fuzzy measures for each subset are obtained. For  $m=3$ , the fuzzy measures are summarized in Table 5.

After all fuzzy measures are identified, and it can be said that the entropy based fuzzy measures are relatively larger than the complexity based fuzzy measures. Furthermore, “not equal thresholds approach” in which the variables can have different thresholds is used. As explained in the methodology section, “histogram function”<sup>†</sup> in Matlab is used as bin width optimization method. When “histogram function” is employed, the threshold numbers of  $D_1, D_2, D_3$ , and  $D_4$  courses were found as 6, 7, 6 and 3, respectively. (For *EKON* course, the threshold number was equal to 9). It is observed that the entropy based fuzzy measures are relatively larger than the complexity based fuzzy measures as seen in Table 6.

<sup>†</sup> The function selects the optimal bin size of a histograms by using automatic binning algorithm such as auto, scott, freedman-diaconis, sturges. These algorithms return bins with a uniform width by showing the underlying shape of the distribution.



**Table 5.** Identified fuzzy measure for  $m = 3$  (Equal thresholds)

| $D_1$ | $D_2$ | $D_3$ | $D_4$ | Entropy based fuzzy measure | Complexity based fuzzy measure |
|-------|-------|-------|-------|-----------------------------|--------------------------------|
| 0     | 0     | 0     | 0     | 0                           | 0                              |
| 1     | 0     | 0     | 0     | 0.37                        | 0.16                           |
| 0     | 1     | 0     | 0     | 0.34                        | 0.16                           |
| 0     | 0     | 1     | 0     | 0.27                        | 0.16                           |
| 0     | 0     | 0     | 1     | 0.27                        | 0.16                           |
| 1     | 1     | 0     | 0     | 0.68                        | 0.42                           |
| 1     | 0     | 1     | 0     | 0.58                        | 0.32                           |
| 1     | 0     | 0     | 1     | 0.60                        | 0.42                           |
| 0     | 1     | 1     | 0     | 0.61                        | 0.42                           |
| 0     | 1     | 0     | 1     | 0.55                        | 0.37                           |
| 0     | 0     | 1     | 1     | 0.53                        | 0.32                           |
| 1     | 1     | 1     | 0     | 0.86                        | 0.68                           |
| 1     | 1     | 0     | 1     | 0.83                        | 0.74                           |
| 1     | 0     | 1     | 1     | 0.73                        | 0.58                           |
| 0     | 1     | 1     | 1     | 0.79                        | 0.63                           |
| 1     | 1     | 1     | 1     | 1                           | 1                              |

**Table 6.** Identified fuzzy measure (Not equal thresholds)

| $D_1$ | $D_2$ | $D_3$ | $D_4$ | Entropy based fuzzy measure | Complexity based fuzzy measure |
|-------|-------|-------|-------|-----------------------------|--------------------------------|
| 0     | 0     | 0     | 0     | 0.00                        | 0.00                           |
| 1     | 0     | 0     | 0     | 0.51                        | 0.21                           |
| 0     | 1     | 0     | 0     | 0.51                        | 0.24                           |
| 0     | 0     | 1     | 0     | 0.42                        | 0.21                           |
| 0     | 0     | 0     | 1     | 0.23                        | 0.10                           |
| 1     | 1     | 0     | 0     | 0.86                        | 0.72                           |
| 1     | 0     | 1     | 0     | 0.83                        | 0.62                           |
| 1     | 0     | 0     | 1     | 0.68                        | 0.45                           |
| 0     | 1     | 1     | 0     | 0.80                        | 0.62                           |
| 0     | 1     | 0     | 1     | 0.66                        | 0.41                           |
| 0     | 0     | 1     | 1     | 0.60                        | 0.34                           |
| 1     | 1     | 1     | 0     | 0.99                        | 0.97                           |
| 1     | 1     | 0     | 1     | 0.91                        | 0.83                           |
| 1     | 0     | 1     | 1     | 0.90                        | 0.76                           |
| 0     | 1     | 1     | 1     | 0.90                        | 0.79                           |
| 1     | 1     | 1     | 1     | 1                           | 1                              |

After the fuzzy measures are identified, the results are intermingled with the discrete Choquet integral through equation (1). By this way the scores of students' academic performances for both the entropy based Choquet integral method and the complexity based Choquet integral method was obtained. When equal thresholds are used, these obtained scores are transformed according to  $m$  level ( $m = 2, 3, 4, 5, 6, 7, 8,$  and  $9$ ) for each entropy based Choquet integral method and complexity based Choquet integral method.

Now, let's consider equal threshold approach. For example, the number of the level of score is equal to 3 (i.e.  $m=3$ ), and the fuzzy measure is entropy based fuzzy measure. The raw scores of the first student are 55.8, 42, 52, 76 (see Table 2). First of all, the scores should be ranked from the smallest to the largest, i.e., 42, 52, 55.8, 76. Then, the estimation score is computed by the discrete Choquet integral as follow:

Estimation score

$$\begin{aligned}
 &= 42 * \mu_1(D_2, D_3, D_1, D_4) + (52 - 42) * \mu_1(D_3, D_1, D_4) + (55.8 - 52) * \mu_1(D_1, D_4) \\
 &\quad + (76 - 55.8) * \mu_1(D_4) \\
 &= 42 * 1.00 + (52 - 42) * 0.73 + (55.8 - 52) * 0.60 + (76 - 55.8) * 0.27 \\
 &= 57.03
 \end{aligned}$$

After all estimation scores of the students' academic performances is computed, the estimation scores are transformed to the categorical data by using "hist.m" program of Matlab. Finally, both the estimation scores and the transformed scores are showed in Table 7 for each students.

**Table 7.** Estimation score and the transformed scores of the students for  $m = 3$

| Student | Estimation score | Transformed score | Student | Estimation score | Transformed score |
|---------|------------------|-------------------|---------|------------------|-------------------|
| 1       | 57.03            | 1                 | 18      | 71.81            | 2                 |
| 2       | 63.78            | 2                 | 19      | 57.33            | 1                 |
| 3       | 47.11            | 1                 | 20      | 71.81            | 2                 |
| 4       | 58.02            | 1                 | 21      | 66.12            | 2                 |
| 5       | 66.37            | 2                 | 22      | 63.93            | 2                 |
| 6       | 72.81            | 2                 | 23      | 78.08            | 3                 |
| 7       | 56.26            | 1                 | 24      | 52.61            | 1                 |
| 8       | 63.31            | 2                 | 25      | 72.73            | 2                 |
| 9       | 63.43            | 2                 | 26      | 60.66            | 2                 |
| 10      | 52.71            | 1                 | 27      | 74.03            | 3                 |
| 11      | 65.35            | 2                 | 28      | 54.89            | 1                 |
| 12      | 71.82            | 2                 | 29      | 70.06            | 2                 |
| 13      | 54.83            | 1                 | 30      | 86.26            | 3                 |
| 14      | 51.70            | 1                 | 31      | 57.42            | 1                 |
| 15      | 50.21            | 1                 | 32      | 76.64            | 3                 |
| 16      | 69.71            | 2                 | 33      | 52.40            | 1                 |
| 17      | 69.17            | 2                 |         |                  |                   |

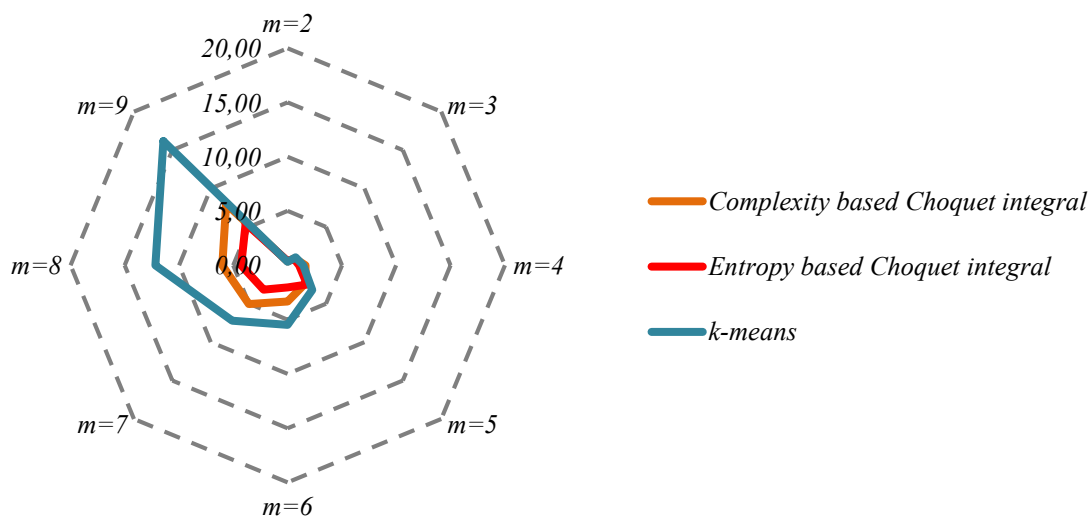
When the complexity based fuzzy measure is used, final transformed scores can be obtained similarly with using the discrete Choquet integral. Now, the evaluation of the performances of each models is required. As explained in section 2.4, mean square errors are used to compare the alternative models performances. In the present study, the *EKON* scores of the students are used as control group, actually these scores are the parameters ( $\theta$  values) and the obtained results by using the alternative methods are the estimators ( $\hat{\theta}$  values). The mean of the squared difference between the parameter and the estimator gives the mean squared error value. The mean square errors are calculated for each method for  $\forall m = 2, 3, 4, 5, 6, 7, 8, 9$ , and the results are shown in Table 8.

**Table 8.** MSE results (Equal threshold)

| $m$   | Complexity based Choquet | Entropy based Choquet |
|-------|--------------------------|-----------------------|
| $m=2$ | 0.36                     | 0.39                  |
| $m=3$ | 0.91                     | 0.85                  |
| $m=4$ | 1.64                     | 1.06                  |
| $m=5$ | 2.33                     | 2.52                  |
| $m=6$ | 3.33                     | 2.03                  |
| $m=7$ | 5.06                     | 3.15                  |
| $m=8$ | 6.03                     | 4.27                  |
| $m=9$ | 8.00                     | 5.48                  |

Finally, the *MSE* results of the methods are summarized in Table 8. Obviously as shown in table, the complexity-based and the entropy-based Choquet integral have the minimum *MSE* results while the number of the level of score ( $m$ ) is two. However, a binary transformation is not generally preferred in the higher institution of learning. By using the idea of this, it can be seen that the complexity-based Choquet integral while  $m = 3, 4, 5$ , and the entropy-based Choquet integral while  $m = 3, 4, 5, 6$  have relatively small *MSE*. Thus,  $m = 3, 4, 5$  can be regarded as possible candidates that should be used in this part of the study. Namely, it can be said that the obtained *MSE* results by using both entropy and complexity based methods are closer to the scores of control group when the number of the level of score is equal to 3, 4 or 5. It is seen that using “equal threshold” Choquet integral both entropy and complexity based provide better results than “not equal threshold” cases in most of the times. The “not equal thresholds” *MSE* results for the entropy and the complexity based Choquet integral are respectively 2.94 and 1.91.

*Robustness Check.* The  $k$ -means is one of the most well-known statistical methods for determining new structure when investigating data sets (Flynt and Dean 2016). The method is widely used for evaluating students’ performances (Veeramuthu et al. 2014). Now, robustness check was provided by comparing  $k$ -means performance with Choquet integral applications. Here the intermediate steps of  $k$ -means algorithm were not provided. (However, if requested, corresponding author can provide the all steps of robustness check using  $k$ -means method).



**Figure 3.** The *MSE* results of the methods

The *MSE* results of  $k$ -means method are respectively 0.33, 1.03, 1.45, 3.21, 5.48, 7.21, 12.18, and 16.19 for  $c = 2, 3, 4, 5, 6, 7, 8$ , and 9.  $k$ -means results can be compared with only “equal threshold” approach results. As the number of the cluster increases, it is seen that the *MSE* value increases. Besides, it can be seen that as the number of the level of score increases, *MSE* value increases. Nevertheless, if “equal threshold” method is used, this increase is less than it is if  $k$ -means method is used. By using the idea of the model with the smallest *MSE* value, the results of the robustness analysis indicate that both entropy and complexity based discrete Choquet integral provides better results than  $k$ -means method in most of the cases as shown in Figure 3.

#### 4. CONCLUSIONS and REMARKS

Evaluation of the academic performance, that takes a wide variety of methods, is an integral part of educational system. That evaluation depends on many criteria that can be seen as a MCDM problem. These problem refers to the analysis and judgment process of selecting an optimal solution from two or more feasible schemes with multiple indicators in order to achieve a certain goal. As for the Choquet integral operator of fuzzy measure, since Schmeidler (1989) first applied it to related MCDM analysis, it has been widely used in decision-making fields for performance evaluation such as engineering, economy and management areas (Xu, 2010; Sun et al., 2015; Han & Wei, 2017; Liu et al., 2018).

At the present time, most of the traditional evaluation techniques take no account of the interactions among criteria. In this regard, the Choquet integral is an effective and appropriate method drawing strong attention to inherently dependent evaluation criteria. In this study, an extensive comparison of several discretization techniques is mapped out for objectively evaluating academic performance of the students. In detail, the discrete Choquet integral is used with the ultimate aim of evaluating the students' success at a university in Turkey. Even though, a specific framework is provided, the method can also be used in any educational assessment such as teacher competency in higher institution of learning and universities perform according to different educational indicators. Thus, the method can be seen as a tool that attracts a good deal of attention in educational assessment.

In this study, the entropy-based and the complexity-based discrete Choquet integral and the  $k$ -means method is used. For the ex-post evaluation, the mean square error method is used in our study. Previous works on the evaluation of students' performance by using the discrete Choquet integral such as Shieh et al., and Chang et al., (2009) did not consider whether the data matrix was normally distributed. However, this study showed that if the data matrix is not normally distributed, entropy-based Choquet integral provides much better results. On the other hand, complexity-based Choquet integral generally presents optimal results if the data is close to being normally distributed. Besides, the other previous studies can show a good performance and a good accuracy results when the sample size is large, but it cannot be possible to deal with the problems when the size is small. Another important aspect of our evaluation is that the paper presents the  $k$ -means method as a robustness analysis to compare the effectiveness of the discrete Choquet integral based methods. The most remarkable property of  $k$ -means is its efficiency in large sample size. However, the obtained mean square error results of the  $k$ -means method indicate that both entropy and complexity based Choquet integral method provides better results than the  $k$ -means method in most of the cases. In conclusion, this study's findings point out that the discrete Choquet integral method provides a major support to educational system in evaluating students' performance.

#### ORCID

Deniz Koçak  <https://orcid.org/0000-0002-5893-0564>

#### 5. REFERENCES



- Angilella, S., Arcidiacono, S.G., Corrente, S., Greco, S., & Matarazzo, B. (2017). An application of the SMAA–Choquet method to evaluate the performance of sailboats in offshore regattas. *Operational Research*, doi:10.1007/s12351-017-0340-7
- Angilella, S., Corrente, S., & Greco, S. (2015). Stochastic multiobjective acceptability analysis for the Choquet integral preference model and the scale construction problem. *European Journal of Operational Research*, 240(1), 172 - 182, doi: 10.1016/j.ejor.2014.06.031
- Baker, R.S.J.D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3-17.

- Branke, J., Corrente, S., Greco, S., Słowiński, R., & Zielniewicz, P. (2016). Using Choquet integral as preference model in interactive evolutionary multiobjective optimization. *European Journal of Operational Research*, 250(3), 884-901, doi: [10.1016/j.ejor.2015.10.027](https://doi.org/10.1016/j.ejor.2015.10.027)
- Calvo, T., Mayor, G., & Mesiar, R. (2002). *Aggregation operators: New trends and applications*. Physica-Verlag, Heidelberg, New York.
- Catlett, J. (1991). *On changing continuous attributes into ordered discrete attributes*. Kodratoff, Y. (Eds.) Machine Learning Lecture Notes in Computer Science, Springer, Berlin, Heidelberg.
- Chang, H.J., Liu, H.C., Tseng, S.W., & Chang, F.M. (2009). A comparison on Choquet integral with different information-based fuzzy measures. In *Proceedings of the 8th International Conference on Machine Learning and Cybernetics* (pp. 3161-3166).
- Chmielewski, M.R., & Grzymala-Busse, J.W. (1996). Global discretization of continuous attributes as preprocessing for machine learning. *International Journal of Approximate Reasoning*, 15(4), 319-331. doi: [10.1016/S0888-613X\(96\)00074-6](https://doi.org/10.1016/S0888-613X(96)00074-6)
- Choquet, G. (1954). Theory of capacities. *Annals of Institute of Fourier*, 5, 131-295.
- Cui, L., & Li, Y. (2008). Linguistic quantifiers based on Choquet integrals. *International Journal of Approximate Reasoning*, 48(2), 559-582. doi: [10.1016/j.ijar.2007.11.001](https://doi.org/10.1016/j.ijar.2007.11.001)
- Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *12th International Conference on Machine Learning*. Los Altos, CA: Morgan Kaufmann, (pp. 194-202).
- Flynt, A., & Dean, N. (2016) A survey of popular R packages for cluster analysis. *Journal of Educational and Behavioral Statistics*, 41(2), 205-225. doi: [10.3102/1076998616631743](https://doi.org/10.3102/1076998616631743)
- Garcia, S., Luengo, J., Saez, A., Lopez, V., & Herrera, F. (2013). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4), 734-750. doi: [10.1109/TKDE.2012.35](https://doi.org/10.1109/TKDE.2012.35)
- Grabisch, M. (1996). The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, 89(3), 445-456, doi: [10.1016/0377-2217\(95\)00176-X](https://doi.org/10.1016/0377-2217(95)00176-X)
- Greene, W.H. (2016). *Econometric Analysis*. 7th Edition, Pearson Education, Inc., Publishing as Prentice Hall. NJ 074458, USA.
- Han, L., & Wei, C. (2017). Group decision making method based on single valued neutrosophic Choquet integral operator. *Operations Research Transactions*, 21(2), doi: [10.15960/j.cnki.issn.1007-6093.2017.02.012](https://doi.org/10.15960/j.cnki.issn.1007-6093.2017.02.012)
- Herde, C.N., Wüstenberg, S., & Greiff, S. (2016). Assessment of complex problem solving: What we know and what we don't know. *Applied Measurement in Education*, 29(4), 265-277, doi: [10.1080/08957347.2016.1209208](https://doi.org/10.1080/08957347.2016.1209208)
- Huber, S.G., & Skedsmo, G. (2016). Teacher evaluation accountability and improving teaching practices. *Educational Assessment, Evaluation and Accountability*, 28(3), 105-109, doi: [10.1007/s11092-016-9241-1](https://doi.org/10.1007/s11092-016-9241-1)
- Jain, A.K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8), 651-666, doi: [10.1016/j.patrec.2009.09.011](https://doi.org/10.1016/j.patrec.2009.09.011)
- Kasparian, J., & Rolland, A. (2012). OECD's 'Better Life Index': Can any country be well ranked?. *Journal of Applied Statistics*, 39(10), 2223 - 2230, doi: [10.1080/02664763.2012.706265](https://doi.org/10.1080/02664763.2012.706265)
- Kononenko, I., & Kukar, M. (2007). *Machine learning and data mining: Introduction to principles and algorithms*. Harwood Publishing Limited.
- Liu, W.F., Du Y. X., & Chang J. (2018). Intuitionistic fuzzy interaction choquet integrals operators and applications in decision making. *Fuzzy systems and Mathematics*, 32(2), doi: [1001-7402\(2018\)02-0110-11](https://doi.org/10.1001-7402(2018)02-0110-11)

- Liu, H., Hussain, F., Tan, C.L., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6(4), 393-423, doi: [10.1023/A:1016304305535](https://doi.org/10.1023/A:1016304305535)
- Marichal, J.L., & Roubens, M. (2000). Determination of weights of interacting criteria from a reference set. *European Journal of Operational Research*, 124(3), 641-650, doi: [10.1016/S0377-2217\(99\)00182-4](https://doi.org/10.1016/S0377-2217(99)00182-4)
- Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 41(4), 1432-1462, doi: [10.1016/j.eswa.2013.08.042](https://doi.org/10.1016/j.eswa.2013.08.042)
- Pötzelberger, K., & Felsenstein, K. (1993). On the fisher information of discretized data. *The Journal of Statistical Computation and Simulation*, 46(3-4), 125-144.
- Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann Publishers, Inc.
- Schmeidler D. (1989). Subjective probability and expected utility without additivity. *Econometrica*, 57(5), 571-587, doi: [10.2307/1911053](https://doi.org/10.2307/1911053)
- Shieh, J.I., Wu, H.H., & Liu, H.C. (2009). Applying a complexity-based Choquet integral to evaluate students' performance. *Expert Systems with Applications*, 36(3), 5100-5106, doi: [10.1016/j.eswa.2008.06.003](https://doi.org/10.1016/j.eswa.2008.06.003)
- Slater, S., Joksimovic, S., Kovanovic, V., & Baker, R.S. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, 42(1), 85-106, doi: [10.1016/j.eswa.2008.06.003](https://doi.org/10.1016/j.eswa.2008.06.003)
- Sun, H. X., Yang, H. X., Wu, J. Z., & Ouyang, Y. (2015). Interval neutrosophic numbers Choquet integral operator for multi-criteria decision making. *Journal of Intelligent & Fuzzy Systems*, 28(6), 2443-2455.
- Tan, P.N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Pearson, Addison Wesley.
- Veeramuthu, P., Periyasamy, R., & Sugasini, V. (2014). Analysis of student result using clustering techniques *International Journal of Computer Science and Information Technologies*, 5(4), 5092-5094.
- Wang, R.-S. & Ha, M.-H. (2008). On the properties of sequences of fuzzy-valued Choquet integrable functions. *Fuzzy Optimization and Decision Making*, 7, 417-431, doi: [10.1007/s10700-008-9040-3](https://doi.org/10.1007/s10700-008-9040-3)
- Wang, Z., Nian, Y., Chu, J., & Shi, Y. (2012). *Emerging Computation and Information Technologies for Education*. Mao, E., Xu, L., & Tian, W. (Eds.), Springer Verlag, Berlin Heidelberg.
- Wen, X., Yan, M., Xian, J., Yue, R. & Peng, A. (2016). Supplier selection in supplier chain management using Choquet integral-based linguistic operators under fuzzy heterogeneous environment. *Fuzzy Optimization and Decision Making*, 15, 307-330.
- Xu, Z. (2010). Choquet integrals of weighted intuitionistic fuzzy information. *Information Sciences*, 180(1), 726-736, doi: [10.1016/j.ins.2009.11.011](https://doi.org/10.1016/j.ins.2009.11.011)
- Zopounidis, C., Doumpos, M. (2002). Multicriteria classification and sorting methods: A literature review. *European Journal of Operational Research*, 138(2), 229-246, doi: [10.1016/S0377-2217\(01\)00243-0](https://doi.org/10.1016/S0377-2217(01)00243-0)



## Improved Performance of Model Fit Indices with Small Sample Sizes in Cognitive Diagnostic Models

Hueying Tzou <sup>1\*</sup>, Ya-Huei Yang <sup>1</sup>

<sup>1</sup> Department of Education, National University of Tainan, Tainan, Taiwan

### ARTICLE HISTORY

Received: 14 November 2018

Revised: 13 March 2019

Accepted: 15 March 2019

### KEYWORDS

RSS,

$\zeta^2$  index,

model fit indices,

cognitive diagnostic models,

Q-matrix

**Abstract:** Selecting an appropriate cognitive diagnostic model (CDM) for data analysis is always challenging. Studies have explored several model fit indices for CDMs. The common results of these studies indicate that Q-matrix misspecifications lead to poor performance of the model fit indices in the context of CDMs. Thus, this study explored whether model fit indices improve performance with a modified Q-matrix. The average class size has reduced to 23 students in Taiwan because of the low birth rate; therefore, the study sought the effect of sample size on the performance of model fit indices. The results showed that Akaike's information criterion (AIC) was an excellent model fit index in small samples. Model fit indices with the modified Q-matrix presented superior performance.

## 1. INTRODUCTION

Recently, cognitive diagnostic models (CDMs) (DiBello, Roussos, & Stout, 2007) have been extensively studied in educational research (Jiao, 2009). CDMs are psychological models that are used to examine whether a subject is proficient in a skill or possesses a particular character (Chen, de la Torre, & Zhang, 2013) in order to provide more precise information regarding the subject (Ma, Iaconangelo, & de la Torre, 2016). When applying CDMs to analyze testing data to obtain diagnostic information regarding a subject, one must select the analytical model and define the Q-matrix of the test (Tatsuoka, 1983). Recently, CDMs have been developed in accordance with their applicable circumstances for different cognitive situations, such as the deterministic inputs, noisy “and” gate model (DINA; Junker & Sijtsma, 2001); the deterministic inputs, noisy “or” gate model (DINO; Templin & Henson, 2006); and the generalized deterministic inputs, noisy “and” gate model (GDINA; de la Torre, 2011). To ensure that valid diagnostic information is obtained from the model analytics, the model–data fit must be considered. Researchers can directly adopt the saturation model for data analysis and routinely have a high degree of fit; however, the complexity of the saturation model requires larger samples to produce accurate estimates (de la Torre & Lee, 2013). Practitioners are often unable to obtain sufficient samples; therefore, the application of CDMs to small sample sizes is critical.

**CONTACT:** Hueying Tzou ✉ [tzou@mail.nutn.edu.tw](mailto:tzou@mail.nutn.edu.tw) 📧 Department of Education, National University of Tainan, 33, Sec. 2, Shu-Lin St., Tainan 700, Taiwan

ISSN-e: 2148-7456 /© IJATE 2019

For example, the average class size has reduced to 23 students in Taiwan because of the low birth rate.

In addition to selecting the correct CDM, the correct Q-matrix is equally critical in CDM analysis. Studies have confirmed that a misspecified Q-matrix negatively affects the recovery of parameters and the classification of subjects (Kunina-Habenicht, Rupp, & Wilhelm, 2012; Rupp & Templin, 2008). Kunina-Habenicht (2012) indicated that with 30% misspecification of the Q-matrix at a sample size of 1000, the accurate classification rate was only 64%, even if the number was increased to 10,000 (10 times) under the same conditions.

Model fit indices were developed to select the appropriate model for data analysis. The indices are mainly divided into two types: absolute and relative fit indices. Common absolute fit indices are pair proportion correct, pair transformed correlation, and pair log-odds. According to Chen (2013), the pair proportion correct rate is a single-variable absolute fit index, and the performance of this indicator is poor; thus, it is not employed in this study. This study explored the performance of pair transformed correlation (hereinafter referred to as  $r$ ) and pair log-odds indices ( $I$ ) to correctly reject wrong models with the modified Q-matrix.

Relative fit indices are another type of model fit indices. More than two models can fit the same data set. To select the most appropriate models, relative model fit indices are required. Relative model fit indices are used for model comparisons and maximum log-likelihood (for example,  $-2 \log$ -likelihood or  $-2LL$ ). The two most commonly used are the Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1976). This study explored the performance of AIC and BIC with the modified Q-matrix.

## 2. BACKGROUND

Studies have noted that Q-matrix misspecification affects model parameters (Rupp & Templin, 2008; de la Torre, 2008) and the accuracy of examinees' classifications (Chiu & Douglas, 2013). If a q-vector of an item was misspecified, the estimated item parameters and the examinees' classifications were significantly biased.

For this reason, researchers have focused on the development of Q-matrix correction methods. de la Torre (2008) developed the sequential  $\delta$  method for the DINA model to perform item-by-attribute Q-matrix modification. According to de la Torre, if an item must be included in a particular attribute, the difference in the correct answering probability of the group with and without the particular attribute is maximized (de la Torre indicated the difference value as  $\delta$ ). Therefore, under the item level, we first assume the q-vector of the item as a zero vector and compare the  $\delta$  values of each attribute (or a combination of attributes) to include the attribute with the maximum  $\delta$  value into the q-vector.

de la Torre simulated 5000 examinees with uniform attribute distribution to explore the performance of the modification method ( $\delta$  method) with different types of Q-matrix misspecification: overspecified (an attribute that is originally not measured but included in the q-vector), underspecified (an attribute that is originally measured but excluded in the q-vector), and mixed misspecification (both overspecified and underspecified in the same q-vector). The results showed that an appropriate cutting value would lead to an excellent modified Q-matrix (same as the original Q-matrix) regardless of the Q-matrix misspecification. However, there are numerous restrictions to this application; the method is only for the DINA model, and the fitting model must be known prior.

Chiu (2013) developed the minimum residual sum of squares (RSS) method to improve the limits of the  $\delta$  method (the fitting model must be known prior). The RSS method is based on nonparametric classification (Chiu & Douglas, 2013) to obtain examinees' attribute patterns and the theoretical response ( $\eta_{ij}$ ) of examinees' attribute patterns and the Q-matrix. The squared

value of the difference between the actual and theoretical responses is calculated with equation (1). Chiu argues that if the q-vector is correctly defined, theoretical responses are similar to actual responses, and the RSS value is minimized. The next step is to calculate the RSS value of each q-vector in the item level and choose the q-vector with the minimum RSS value as the new q-vector of the corresponding item.

$$RSS_i = \sum_{j=1}^N (X_{ij} - \eta_{ij})^2 \quad (1)$$

There were two data-generating models (DINA and noisy input, deterministic “and” gate [NIDA]), two attribute numbers ( $K = 3$  and  $5$ ), three attribute pattern distributions (uniform, multivariate normal threshold, and higher order), three sample sizes (100, 500, and 1000), four item qualities ( $s = g = 0.2, 0.3, 0.4, \text{ and } 0.5$ ), and two Q-matrix misspecification rates (random misspecification 10% or 20%). The criteria were to compare the recovery rates of the true and corrected Q-matrices. A higher recovery rate indicated the superior correction performance of the RSS method. The results showed that despite the small sample size, in the case of 0.3, the recovery rate was at least 88% if the Q-matrix misspecification was 10% and at least 75% if the Q-matrix misspecification was 20%.

Unlike the previous  $\delta$  index which is only applied for the DINA model and required assuming about the fitting model, de la Torre and Chiu (2016) developed another more generalized Q-matrix modification method; they called this index  $\zeta^2$  (de la Torre & Chiu, 2016). The new modification method  $-\zeta^2$ , used the GDINA model to exceed the limits of the  $\delta$  method, which was only applied with the DINA model. However, the GDINA model is a complex model because of the estimation of many parameters. In other words, a large sample size is required to obtain accurate estimates. The sample size in the study of de la Torre & Chiu (2016) was 2000; the performance of smaller samples has rarely been explored. Therefore, this study mainly focuses on small sample sizes and explores the performance of  $\zeta^2$  indicators.

As showed in Chiu’s (2013) study, the performance of the RSS method with the data generated from the DINA and NIDA models was excellent with small sample sizes. Nevertheless, the performance of the RSS method under the GDINA model was rarely discussed in literatures. In this case, we compare the performance of the RSS method and the  $\zeta^2$  method under the GDINA model.

### 3. METHOD

#### 3.1. Research Purposes and Questions

The purposes of the study are as follows:

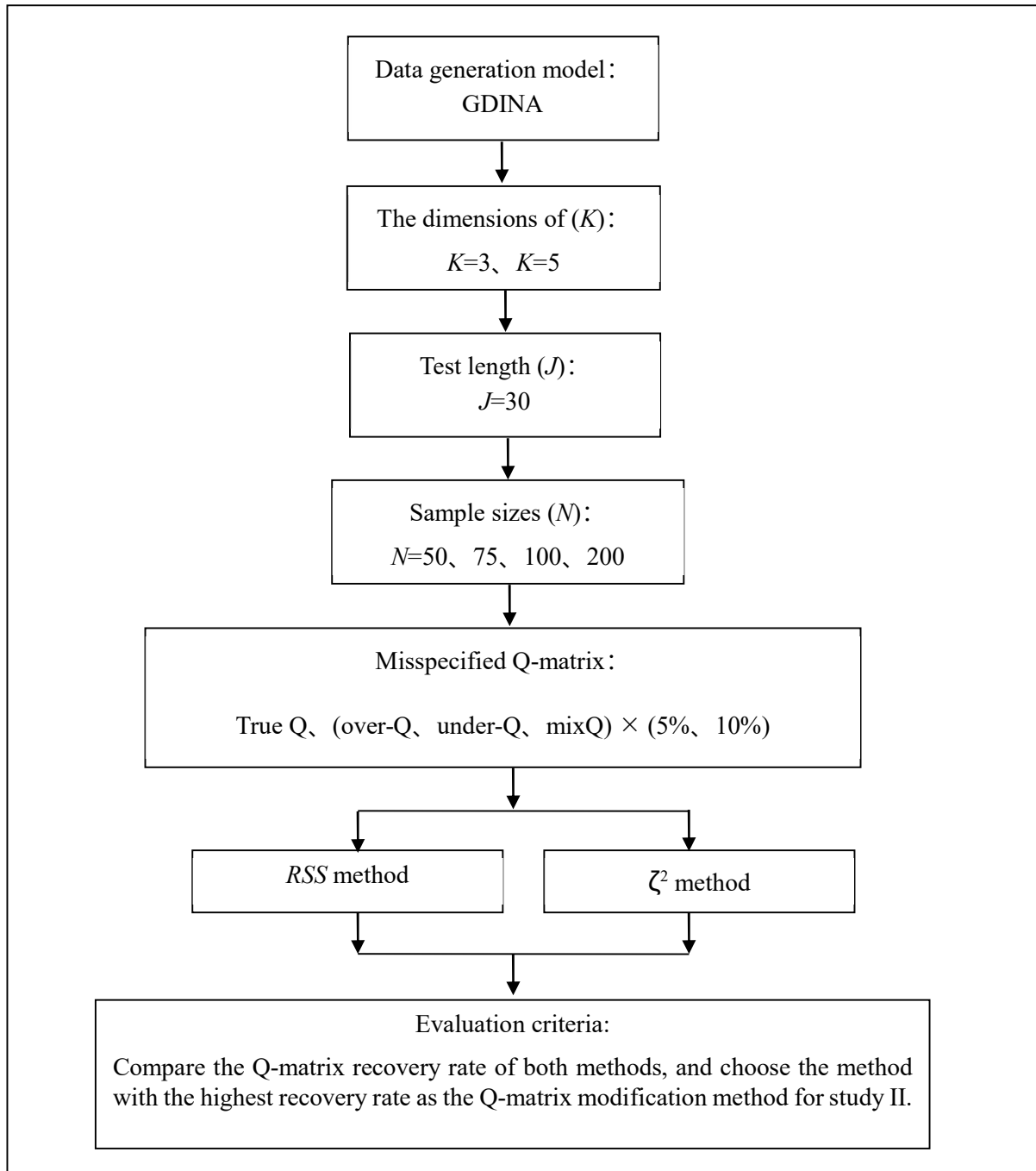
1. To explore the performance of the RSS method and the  $\zeta^2$  method with the setting sample sizes and Q-matrix misspecifications.
2. To explore the performance of model fit indices (AIC, BIC,  $r$ ,  $l$ ) with small sample sizes with the original and modified Q-matrix.
3. To compare the performance of the model fit indices with the original and modified Q-matrix.

#### 3.2. Study I: Simulation design

This research was divided into two studies. In study I, both Q-matrix modification methods were compared, and the superior one would be used in the second study. Figure 1 presents the flow chart of study I.

Data generations and analyses were conducted using R software (R Core Team, 2017). The R package GDINA (Ma & de la Torre, 2018) was used to generate data sets. The item parameters were setting as  $s=g=0.1$  for all items. We assumed the examinees’ attribute patterns were uniform. The number of attributes varied with the coverage of test; thus, we assumed that the smaller domain contained fewer attributes ( $K = 3$ ), and the larger domain contained more

attributes ( $K = 5$ ). Test length was set to 30 items. Meanwhile, Taiwan currently averages 23 students per class and 3.3 classes per grade in elementary schools. Therefore, we set the sample sizes to 50 (approximately two classes), 75 (three classes), 100, and 200.



**Figure1.** Experimental flow chart for study I.

A total of six Q-matrix misspecification situations (three misspecification types  $\times$  two misspecification rates) were studied; the three misspecification types were overspecified (overQ), underspecified (underQ), and mix-specified (mixQ) Q-matrices. The misspecifications were randomly altered. OverQ meant that the item did not require the attribute, but the coding of the attribute was changed to 1 from 0 to become a requiring attribute; underQ meant that the item required the attribute, but the coding of the attribute was altered to 0; mixQ meant that in the same item, one required attribute was coded as 0 and another one not

required attribute was coded as 1. Misspecification rates were 5% and 10%. For example, 5%overQ meant 5% erroneous coding elements overspecified in the Q-matrix

In this study, we compared the Q-matrix recovery rates of both modification methods (RSS method and  $\zeta^2$  method) with the six Q-matrix misspecifications and eight simulation conditions. We simulated each of the 48 combinations in this study and replicated each conditions 30 times. The results were displayed with the mean Q-matrix recovery rate.

$$Q\text{-matrix recovery rates} = 1 - \left| \frac{\sum_{j=1}^J \sum_{k=1}^K q_{jk}^{\text{original}} - q_{jk}^{\text{corrected}}}{J \times K} \right| \tag{2}$$

J: test length

K: numbers of attributes

$q_{jk}^{\text{original}}$  : original coding in item j and attribute k

$q_{jk}^{\text{corrected}}$  : corrected coding in item j and attribute k

Many researches have shown that a misspecified Q-matrix affects the estimation of item parameters (de la Torre, 2008; Rupp & Templin, 2008; Kuninan-Habenicht et al., 2012). To prevent confounding effects on the study results caused by the structure of the Q-matrix, we made the Q-matrix as balanced as possible. The balanced design maintained the number of attributes measured by an item (mean item complexity) and the number of items measuring each attribute (attribute information) approximately the same.

Table 1 shows the correct Q-matrix of K = 5 (hereinafter referred to as True Q, TQ). The attribute information is the same (each attribute is measured by 12 items). There are 10 single-attribute items, 10 double-attribute items, and 10 triple-attribute items. Table 2 is the TQ of K = 3.

**Table 1.** True Q-matrix for K = 5

|        | $\alpha1$ | $\alpha2$ | $\alpha3$ | $\alpha4$ | $\alpha5$ |        | $\alpha1$ | $\alpha2$ | $\alpha3$ | $\alpha4$ | $\alpha5$ |
|--------|-----------|-----------|-----------|-----------|-----------|--------|-----------|-----------|-----------|-----------|-----------|
| Item01 | 1         | 0         | 0         | 0         | 0         | Item16 | 0         | 1         | 0         | 1         | 0         |
| Item02 | 0         | 1         | 0         | 0         | 0         | Item17 | 0         | 1         | 0         | 0         | 1         |
| Item03 | 0         | 0         | 1         | 0         | 0         | Item18 | 0         | 0         | 1         | 1         | 0         |
| Item04 | 0         | 0         | 0         | 1         | 0         | Item19 | 0         | 0         | 1         | 0         | 1         |
| Item05 | 0         | 0         | 0         | 0         | 1         | Item20 | 0         | 0         | 0         | 1         | 1         |
| Item06 | 1         | 0         | 0         | 0         | 0         | Item21 | 1         | 1         | 1         | 0         | 0         |
| Item07 | 0         | 1         | 0         | 0         | 0         | Item22 | 1         | 1         | 0         | 1         | 0         |
| Item08 | 0         | 0         | 1         | 0         | 0         | Item23 | 1         | 1         | 0         | 0         | 1         |
| Item09 | 0         | 0         | 0         | 1         | 0         | Item24 | 1         | 0         | 1         | 1         | 0         |
| Item10 | 0         | 0         | 0         | 0         | 1         | Item25 | 1         | 0         | 1         | 0         | 1         |
| Item11 | 1         | 1         | 0         | 0         | 0         | Item26 | 1         | 0         | 0         | 1         | 1         |
| Item12 | 1         | 0         | 1         | 0         | 0         | Item27 | 0         | 1         | 1         | 1         | 0         |
| Item13 | 1         | 0         | 0         | 1         | 0         | Item28 | 0         | 1         | 1         | 0         | 1         |
| Item14 | 1         | 0         | 0         | 0         | 1         | Item29 | 0         | 1         | 0         | 1         | 1         |
| Item15 | 0         | 1         | 1         | 0         | 0         | Item30 | 0         | 0         | 1         | 1         | 1         |

**Table 2.** True Q-matrix for K = 3

|        | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ |        | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ |
|--------|------------|------------|------------|--------|------------|------------|------------|
| Item01 | 1          | 0          | 0          | Item16 | 0          | 1          | 0          |
| Item02 | 0          | 1          | 0          | Item17 | 0          | 0          | 1          |
| Item03 | 0          | 0          | 1          | Item18 | 1          | 1          | 0          |
| Item04 | 1          | 1          | 0          | Item19 | 1          | 0          | 1          |
| Item05 | 1          | 0          | 1          | Item20 | 0          | 1          | 1          |
| Item06 | 0          | 1          | 1          | Item21 | 1          | 1          | 1          |
| Item07 | 1          | 1          | 1          | Item22 | 1          | 0          | 0          |
| Item08 | 1          | 0          | 0          | Item23 | 0          | 1          | 0          |
| Item09 | 0          | 1          | 0          | Item24 | 0          | 0          | 1          |
| Item10 | 0          | 0          | 1          | Item25 | 1          | 1          | 0          |
| Item11 | 1          | 1          | 0          | Item26 | 1          | 0          | 1          |
| Item12 | 1          | 0          | 1          | Item27 | 0          | 1          | 1          |
| Item13 | 0          | 1          | 1          | Item28 | 1          | 1          | 1          |
| Item14 | 1          | 1          | 1          | Item29 | 1          | 1          | 1          |
| Item15 | 1          | 0          | 0          | Item30 | 1          | 1          | 1          |

### 3.2.1. Q-matrix misspecification design

In the case of 5%overQ, there were 150 elements with 5% misspecifications. The researcher randomly selected 7 elements that originally coded as 0 and altered them to 1. In the case of 5%underQ, 8 elements that originally coded as 1 were altered to 0. In the case of 5%mixQ, 7 items were selected; the elements that originally coded as 0 were altered to 1, and the elements that originally coded as 1 were altered to 0 under the same item. In the case of K = 3, 4 elements were altered in 5%overQ, 5 elements were altered in 5%underQ, and 4 items were altered in 5%mixQ.

### 3.3. Results of Study I: Performance of Q-matrix Modification Methods

The performance of the  $\zeta^2$  index under the condition of K = 3 is shown in Table 3. The lowest recovery rate (0.815) was shown in N = 50; this indicates 18.5% type I error. The highest recovery rate (0.956) was shown in N = 200; this indicates 4.4% type I error. The lowest recovery rate under the condition of K = 5 (0.516) was shown in N = 50; this indicates 48.4% type I error. The highest recovery rate (0.620) was shown in N = 200; this indicates 38% type I error. De la Torre and Chiu (2016) used the  $\zeta^2$  index to modify the data generated from the GDINA model with fixed sample size (N = 2000), test length (J = 30), attribute numbers (K = 5), and random Q-matrix misspecification rates (5%). The result showed that type I error was 2%, and the Q-matrix recovery rate was 0.971. By comparison, the results of the current study showed much higher type I error and a much lower Q-matrix recovery rate. The effect of sample size on the performance of the  $\zeta^2$  method was notable. The results of the current study were quite different from those of de la Torre and Chiu (2016) under the same K = 5 simulation conditions. This may be caused by the sample size. The  $\zeta^2$  method requires item parameters and examinees' attribute patterns for Q-matrix modification. The largest sample size in this study (N = 200) was only one tenth of that in the study of de la Torre and Chiu; therefore, the estimated parameters were less accurate and led to poor modification results. This can be partially supported by the simulation result of K = 3. Because of the decreased number of attributes (K = 3), the estimated parameters were decreased, and the accuracy of parameter estimation was improved. Therefore, the modification performance of the  $\zeta^2$  index method was improved. Despite the sample size of only 200, type I error was reduced to below 5%, and the Q-matrix recovery rate was increased to 0.95.



The results of the RSS method in the case of  $K = 3$  are shown in Table 4. The recovery rate of RSS for TQ was approximately 0.77, and the type I error was approximately 23%. Under  $K = 5$ , the lowest recovery rate (0.812) was shown in  $N = 50$ ; this indicates that the type I error was 18.8%. The highest recovery rate (0.835) was shown in  $N = 200$ ; this indicates that the type I error was 16.5%. The  $\zeta^2$  method exhibited lower type I error in  $K = 3$ ; the RSS method exhibited lower type I error in  $K = 5$ .

For  $K = 3$ , 5% Q-matrix misspecification, and  $N = 200$ , the recovery rate of the  $\zeta^2$  method exceeded 0.95; that is, the misspecification rates of the modified Q-matrix were lower than 5%. In the case of 10% Q-matrix misspecification and  $N = 100$ , the recovery rate of the  $\zeta^2$  method exceeded 0.9. According to the results of  $K = 3$  and  $N = 200$ , the misspecification rate of the modified Q-matrix was lower than that of the original Q-matrix. This indicates that the  $\zeta^2$  method is an effective modification method. However, in the case of  $K = 5$ , the performance of the  $\zeta^2$  method was not acceptable. For  $K = 5$ , the Q-matrix recovery rates were all lower than 0.7; that is, the misspecification rate of the modified Q-matrix was higher than the settings. Therefore, the  $\zeta^2$  method is not suitable for  $K = 5$  and  $N < 200$ .

**Table 3.** Q-Matrix Recovery Rates of the  $\zeta^2$  Method

| K=3 | TQ    | overQ |       | underQ |       | mixQ  |       |
|-----|-------|-------|-------|--------|-------|-------|-------|
|     |       | 5%    | 10%   | 5%     | 10%   | 5%    | 10%   |
| N   |       |       |       |        |       |       |       |
| 50  | 0.815 | 0.817 | 0.813 | 0.816  | 0.820 | 0.810 | 0.801 |
| 75  | 0.865 | 0.864 | 0.860 | 0.864  | 0.866 | 0.865 | 0.852 |
| 100 | 0.902 | 0.899 | 0.900 | 0.902  | 0.900 | 0.897 | 0.901 |
| 200 | 0.956 | 0.953 | 0.952 | 0.954  | 0.952 | 0.954 | 0.952 |
| K=5 | TQ    | overQ |       | underQ |       | mixQ  |       |
| N   |       | 5%    | 10%   | 5%     | 10%   | 5%    | 10%   |
| 50  | 0.516 | 0.511 | 0.504 | 0.531  | 0.546 | 0.525 | 0.538 |
| 75  | 0.512 | 0.499 | 0.498 | 0.522  | 0.549 | 0.522 | 0.542 |
| 100 | 0.516 | 0.510 | 0.506 | 0.536  | 0.550 | 0.541 | 0.550 |
| 200 | 0.620 | 0.608 | 0.596 | 0.638  | 0.653 | 0.631 | 0.637 |

Note:  $N$  = sample size; TQ = True Q-matrix, the Q-matrix used for data generation; overQ = overspecified Q-matrix; underQ = underspecified Q-matrix; mixQ = mix-misspecified Q-matrix; 5% = 5% of entries of the Q-matrix were changed; 10% = 10% of entries of the Q-matrix were changed.

**Table 4.** Q-Matrix Recovery Rates of the RSS Method

| K=3 | TQ    | overQ |       | underQ |       | mixQ  |       |
|-----|-------|-------|-------|--------|-------|-------|-------|
|     |       | 5%    | 10%   | 5%     | 10%   | 5%    | 10%   |
| N   |       |       |       |        |       |       |       |
| 50  | 0.778 | 0.776 | 0.776 | 0.777  | 0.778 | 0.778 | 0.769 |
| 75  | 0.773 | 0.771 | 0.769 | 0.772  | 0.774 | 0.774 | 0.773 |
| 100 | 0.773 | 0.773 | 0.773 | 0.773  | 0.774 | 0.774 | 0.771 |
| 200 | 0.776 | 0.777 | 0.775 | 0.774  | 0.774 | 0.776 | 0.774 |
| K=5 | TQ    | overQ |       | underQ |       | mixQ  |       |
| N   |       | 5%    | 10%   | 5%     | 10%   | 5%    | 10%   |
| 50  | 0.812 | 0.805 | 0.788 | 0.816  | 0.809 | 0.797 | 0.757 |
| 75  | 0.822 | 0.819 | 0.793 | 0.822  | 0.824 | 0.812 | 0.777 |
| 100 | 0.829 | 0.826 | 0.821 | 0.823  | 0.825 | 0.816 | 0.782 |
| 200 | 0.835 | 0.836 | 0.833 | 0.833  | 0.831 | 0.833 | 0.816 |

### 3.3.2. The effect of sample size

In the case of  $K = 3$ , the Q-matrix recovery rates of the  $\zeta^2$  method increased with the sample size; by contrast, the recovery rates of the RSS method were fixed at approximately 0.77, and no increasing trend was observed. Additionally, under the condition of  $K = 3$ , the Q-matrix recovery rates of the  $\zeta^2$  method were higher than those of the RSS method; and the difference in recovery rates between the  $\zeta^2$  and RSS methods increased with sample size.

However, under the condition of  $K = 5$ , the performance of the methods was considerably different. The Q-matrix recovery rates of the RSS method were significantly higher than those of the  $\zeta^2$  method. The difference in both methods was the largest at  $N = 50$  and the smallest at  $N = 200$ . In other words, with larger sample sizes, the recovery rates became more similar.

### 3.3.3. The effect of Q-matrix misspecification rates

Under the condition of  $K = 3$ , the Q-matrix recovery rates of both methods did not reduce with the increase in misspecification rates. For example, the recovery rate of the  $\zeta^2$  method was 0.817 at 5%overQ and 0.813 at 10%overQ. The recovery rates were almost the same even though the misspecification rate increased from 5% to 10%. Furthermore, the recovery rate of the RSS method was 0.776 at 5%overQ and 10%overQ with no difference between the misspecification rates. In the case of  $K = 5$ , the Q-matrix recovery rates of the  $\zeta^2$  and RSS methods decreased slightly due to the increase in misspecification rates, but it is was not significant

### 3.3.4. The effect of Q-matrix misspecification types

The difference in the recovery rates was not distinct among the three Q-matrix misspecification types for the  $\zeta^2$  and RSS methods. For example, in the case of  $K = 3$ ,  $N = 200$ , and 10% misspecification, the recovery rates of the  $\zeta^2$  method were 0.952 for overQ, underQ, and mixQ; meanwhile, the recovery rates of the RSS method were 0.775, 0.774, and 0.774, respectively. The results imply that the Q-matrix misspecification type has a minor effect on both methods.

For  $K = 3$ , the  $\zeta^2$  method exhibited superior modification performance; by contrast, the RSS method exhibited superior modification performance for  $K = 5$ . Given these results, the  $\zeta^2$  method was applied to Q-matrix modification for  $K = 3$ , and the RSS method was applied for  $K = 5$  in study II.

## 3.4. Study II: Simulation design

The simulation data are the same as those used for study I. Study II compared the performance of model fit indices with the true Q-matrix, misspecification Q-matrix, and the corresponding modified Q-matrix. Figure 2 presents the flow chart of study II.

## 3.5. Results of Study II

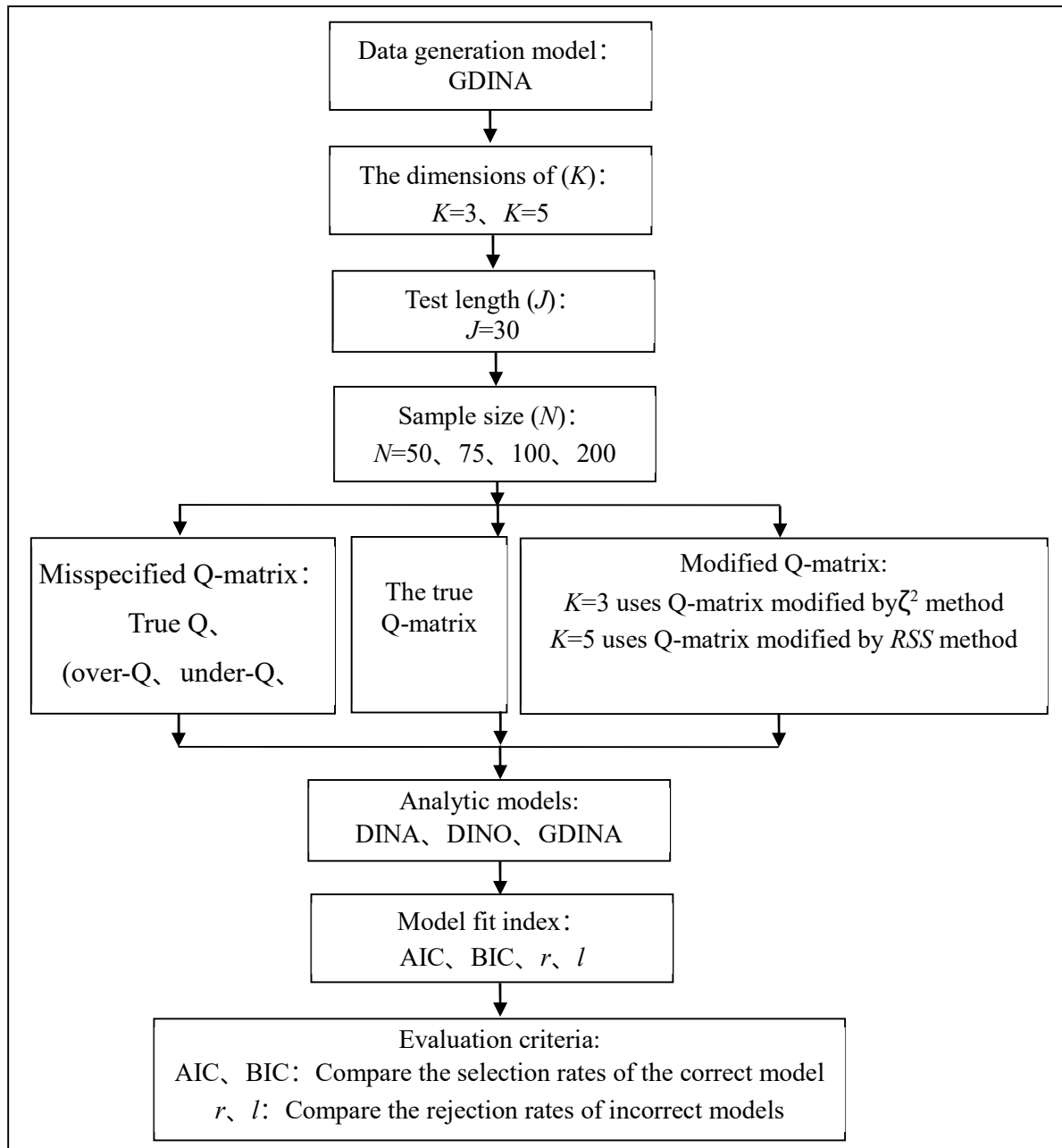
For the readability, the results of  $K=5$  were shown in the appendix page.

### 3.5.1. The performance of relative indices with original Q-matrix

As shown in Table 5, under the condition of  $K = 3$ , AIC always correctly selected the GDINA (the correct data-generating model) as the fitting model; the performance of BIC in correct model selection varied with the misspecification type. In the cases of overQ and underQ, the selection rate of GDINA was much higher than that of the other two models in BIC, except for underQ at  $N = 50$ ; in the case of 10%mixQ, the selection rate of GDINA was much higher than that of the other two models at  $N = 100$  and  $N = 200$ . The results implied that in the case of  $K = 3$ , BIC was considerably affected by Q-matrix misspecification types.

Under the condition of  $K = 5$ , the selection rate of GDINA in AIC was much higher than that of the other two models for all simulation conditions. Different from that in the  $K = 3$  scenario,

the performance of model selection in BIC was affected by sample size for  $K = 5$ . Only for  $N = 200$  was the selection rate of GDINA in BIC much higher than that of the other two models. The results showed that the correct model selection rate of AIC was high under various conditions. BIC was affected by the Q-matrix misspecification type under the condition of  $K = 3$  and by sample size under the condition of  $K = 5$ . These results were similar to the results of Hu et al. (2015).



**Figure 2.** Experimental flow chart for study II.

**Table 5.** Selection Rates of the Relative Indices Under Various Simulation Conditions.

| K=3 |       | TQ  |       | overQ |       |     |       | underQ |       |     |       | mixQ |       |     |       |
|-----|-------|-----|-------|-------|-------|-----|-------|--------|-------|-----|-------|------|-------|-----|-------|
|     |       |     |       | 5%    |       | 10% |       | 5%     |       | 10% |       | 5%   |       | 10% |       |
| N   | M     | AIC | BIC   | AIC   | BIC   | AIC | BIC   | AIC    | BIC   | AIC | BIC   | AIC  | BIC   | AIC | BIC   |
| 50  | DINA  | 0   | 0.167 | 0     | 0.033 | 0   | 0     | 0      | 0.267 | 0   | 0.267 | 0    | 0.333 | 0   | 0.333 |
|     | DINO  | 0   | 0.333 | 0     | 0.100 | 0   | 0.033 | 0      | 0.300 | 0   | 0.267 | 0    | 0.400 | 0   | 0.600 |
|     | GDINA | 1   | 0.500 | 1     | 0.867 | 1   | 0.967 | 1      | 0.433 | 1   | 0.467 | 1    | 0.267 | 1   | 0.067 |
| 75  | DINA  | 0   | 0     | 0     | 0     | 0   | 0     | 0      | 0.033 | 0   | 0.067 | 0    | 0.067 | 0   | 0.167 |
|     | DINO  | 0   | 0.033 | 0     | 0     | 0   | 0     | 0      | 0.133 | 0   | 0.067 | 0    | 0.167 | 0   | 0.467 |
|     | GDINA | 1   | 0.967 | 1     | 1     | 1   | 1     | 1      | 0.833 | 1   | 0.867 | 1    | 0.767 | 1   | 0.367 |
| 100 | DINA  | 0   | 0     | 0     | 0     | 0   | 0     | 0      | 0     | 0   | 0     | 0    | 0     | 0   | 0.033 |
|     | DINO  | 0   | 0     | 0     | 0     | 0   | 0     | 0      | 0     | 0   | 0     | 0    | 0.067 | 0   | 0.200 |
|     | GDINA | 1   | 1     | 1     | 1     | 1   | 1     | 1      | 1     | 1   | 1     | 1    | 0.933 | 1   | 0.767 |
| 200 | DINA  | 0   | 0     | 0     | 0     | 0   | 0     | 0      | 0     | 0   | 0     | 0    | 0     | 0   | 0     |
|     | DINO  | 0   | 0     | 0     | 0     | 0   | 0     | 0      | 0     | 0   | 0     | 0    | 0     | 0   | 0     |
|     | GDINA | 1   | 1     | 1     | 1     | 1   | 1     | 1      | 1     | 1   | 1     | 1    | 1     | 1   | 1     |

Note: M = analytic model

**3.5.2. The performance of relative indices with the modified Q-matrix**

As shown in Table 6, under the condition of K = 3, AIC still correctly selected the GDINA model as the fitting model. The correct model selection rates of BIC with the modified Q-matrix were higher than those with the original Q-matrix. For example, in the case of the original underQ, a perfect correct selection rate of the GDINA in BIC was observed at N = 100 and N = 200 but observed at N = 75, 100, and 200 in the corresponding modified underQ. Under the situation of K = 5, AIC showed better correct selection rates at sample sizes equal to or greater than 75; while BIC showed better correct selection rates at sample sizes equal to 200.

**Table 6.** Selection Rates for Relative Indices with the Modified Q-Matrix

| K=3, ZETA |       | overQ |       |     |       | underQ |       |     |       | mixQ |       |     |       |
|-----------|-------|-------|-------|-----|-------|--------|-------|-----|-------|------|-------|-----|-------|
|           |       | 5%    |       | 10% |       | 5%     |       | 10% |       | 5%   |       | 10% |       |
| N         | M     | AIC   | BIC   | AIC | BIC   | AIC    | BIC   | AIC | BIC   | AIC  | BIC   | AIC | BIC   |
| 50        | DINA  | 0     | 0.133 | 0   | 0.133 | 0      | 0.167 | 0   | 0.133 | 0    | 0.200 | 0   | 0.200 |
|           | DINO  | 0     | 0.067 | 0   | 0.067 | 0      | 0.067 | 0   | 0.067 | 0    | 0.067 | 0   | 0.067 |
|           | GDINA | 1     | 0.800 | 1   | 0.800 | 1      | 0.767 | 1   | 0.800 | 1    | 0.733 | 1   | 0.733 |
| 75        | DINA  | 0     | 0     | 0   | 0     | 0      | 0     | 0   | 0     | 0    | 0     | 0   | 0     |
|           | DINO  | 0     | 0     | 0   | 0     | 0      | 0     | 0   | 0     | 0    | 0     | 0   | 0     |
|           | GDINA | 1     | 1     | 1   | 1     | 1      | 1     | 1   | 1     | 1    | 1     | 1   | 1     |
| 100       | DINA  | 0     | 0     | 0   | 0     | 0      | 0     | 0   | 0     | 0    | 0     | 0   | 0     |
|           | DINO  | 0     | 0     | 0   | 0     | 0      | 0     | 0   | 0     | 0    | 0     | 0   | 0     |
|           | GDINA | 1     | 1     | 1   | 1     | 1      | 1     | 1   | 1     | 1    | 1     | 1   | 1     |
| 200       | DINA  | 0     | 0     | 0   | 0     | 0      | 0     | 0   | 0     | 0    | 0     | 0   | 0     |
|           | DINO  | 0     | 0     | 0   | 0     | 0      | 0     | 0   | 0     | 0    | 0     | 0   | 0     |
|           | GDINA | 1     | 1     | 1   | 1     | 1      | 1     | 1   | 1     | 1    | 1     | 1   | 1     |

Note: M = analytic model

**3.5.3. The performance of absolute indices with the original Q-matrix**

According to Table 7, under the condition of  $K = 3$ , the rejection rates of three models (DINA, DINO, and GDINA) for mixQ were all 1 in both absolute model fit indices ( $r, l$ ); the rejection rates of the three models increased with sample sizes in both absolute model fit indices under the situation of underQ. In the cases of TQ and overQ, the rejection rates of the DINA and DINO models increased with sample size, but the GDINA model decreased with sample size. These results implied that  $r$  and  $l$  tended to reject correct models for underQ and mixQ but tended to fail to reject correct models in the cases of TQ and overQ. These results were consistent with those of Chen et al., (2013) and Hu et al., (2015). Overall, when applying  $r$  and  $l$ , they accepted the correct model under the conditions of TQ and overQ. The results of  $K = 5$  were roughly similar to that of  $K = 3$ .

**Table 7.** Rejection Rates of the Absolute Indices Under Various Simulation Conditions

| K=3 |       | TQ    |       | overQ |       |       |       | underQ |       |       |     | mixQ |     |     |     |
|-----|-------|-------|-------|-------|-------|-------|-------|--------|-------|-------|-----|------|-----|-----|-----|
|     |       |       |       | 5%    |       | 10%   |       | 5%     |       | 10%   |     | 5%   |     | 10% |     |
| N   | M     | $r$   | $l$   | $r$   | $l$   | $r$   | $l$   | $r$    | $l$   | $r$   | $l$ | $r$  | $l$ | $r$ | $l$ |
| 50  | DINA  | 0.900 | 0.900 | 1     | 1     | 1     | 1     | 0.867  | 0.933 | 0.967 | 1   | 1    | 1   | 1   | 1   |
|     | DINO  | 0.800 | 0.900 | 1     | 1     | 1     | 1     | 0.800  | 0.900 | 1     | 1   | 1    | 1   | 1   | 1   |
|     | GDINA | 0.067 | 0     | 0.067 | 0.033 | 0.067 | 0     | 0.533  | 0.567 | 0.867 | 1   | 1    | 1   | 1   | 1   |
| 75  | DINA  | 1     | 1     | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1   | 1    | 1   | 1   | 1   |
|     | DINO  | 1     | 1     | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1   | 1    | 1   | 1   | 1   |
|     | GDINA | 0     | 0.033 | 0     | 0.033 | 0     | 0.033 | 0.667  | 0.667 | 1     | 1   | 1    | 1   | 1   | 1   |
| 100 | DINA  | 1     | 1     | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1   | 1    | 1   | 1   | 1   |
|     | DINO  | 1     | 1     | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1   | 1    | 1   | 1   | 1   |
|     | GDINA | 0     | 0     | 0     | 0     | 0     | 0     | 0.767  | 0.833 | 1     | 1   | 1    | 1   | 1   | 1   |
| 200 | DINA  | 1     | 1     | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1   | 1    | 1   | 1   | 1   |
|     | DINO  | 1     | 1     | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1   | 1    | 1   | 1   | 1   |
|     | GDINA | 0     | 0     | 0     | 0     | 0     | 0     | 0.967  | 0.967 | 1     | 1   | 1    | 1   | 1   | 1   |

**3.5.4. The performance of absolute indices with the modified Q-matrix**

As shown in Table 8, under the condition of  $K = 3$ ,  $r$  and  $l$  tended to fail to reject the GDINA (rejection rates were less than 0.1), and the rejection rates of the DINA and the DINO were all 1. These results showed that the rejection rates of the correct model could be effectively reduced after the Q-matrix was modified. However, there was no such finding for  $K = 5$ . It meant that the rejection rates of the correct model weren't affected by the modified Q-matrix as  $K = 5$ .

**Table 8.** Rejection Rates for Absolute Indices with the Modified Q-Matrix

| K=3, ZETA |       | overQ |       |       |       | underQ |       |       |       | mixQ  |       |       |       |
|-----------|-------|-------|-------|-------|-------|--------|-------|-------|-------|-------|-------|-------|-------|
|           |       | 5%    |       | 10%   |       | 5%     |       | 10%   |       | 5%    |       | 10%   |       |
| N         | M     | r     | l     | r     | l     | r      | l     | r     | l     | r     | l     | r     | l     |
| 50        | DINA  | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1     | 1     | 1     | 1     | 1     |
|           | DINO  | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1     | 1     | 1     | 1     | 1     |
|           | GDINA | 0.033 | 0     | 0.033 | 0     | 0.067  | 0     | 0.067 | 0     | 0.033 | 0     | 0.067 | 0     |
| 75        | DINA  | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1     | 1     | 1     | 1     | 1     |
|           | DINO  | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1     | 1     | 1     | 1     | 1     |
|           | GDINA | 0     | 0     | 0     | 0     | 0.033  | 0.033 | 0.033 | 0.033 | 0     | 0     | 0     | 0     |
| 100       | DINA  | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1     | 1     | 1     | 1     | 1     |
|           | DINO  | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1     | 1     | 1     | 1     | 1     |
|           | GDINA | 0.033 | 0.033 | 0.033 | 0.033 | 0.033  | 0.033 | 0.033 | 0.033 | 0.033 | 0.033 | 0.033 | 0.033 |
| 200       | DINA  | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1     | 1     | 1     | 1     | 1     |
|           | DINO  | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1     | 1     | 1     | 1     | 1     |
|           | GDINA | 0.033 | 0.033 | 0.033 | 0.033 | 0.033  | 0.033 | 0.033 | 0.033 | 0.033 | 0.033 | 0.033 | 0.033 |

#### 4. DISCUSSION

Before applying CDM results, one must ensure the CDM fits the data. Enhancing the fitness of the data and model strengthens the validity and inferences of the results. Therefore, selecting an appropriate model–fit index is crucial. The decreasing birth rate in Taiwan causes the number of students to decrease each year; the teaching and learning style are demanding to provide more individualized information. Conventional single scores have been inappropriate to help teachers, parents, and students to understand learning results. CDMs can exactly meet the needs of the current education and give individual students feedback on learning strengths and weaknesses. Previous studies have shown that misspecified Q-matrices and model selection affect the performance and applicability of CDMs, but no study has designed to explore whether the effect of CDMs applications can be improved as the Q-matrix has been modified in advance. The current study not only explore the effect of CDMs applications with the modified Q-matrix, but also explore the effect of CDMs applications with the small sample sizes to meet the education field needs.

The  $\zeta^2$  index and the RSS methods of Q-matrix modification were explored in this study. According to the modification results, in the case of  $K = 3$ , the  $\zeta^2$  method can effectively correct the misspecification of the Q-matrix. However, in the case of  $K = 5$ , the performance of both methods was not as good as expected. We also found that the performance RSS method was affected by the data generation models. In this study, for  $K = 3$ , the Q-matrix recovery rates of the RSS method were in the interval of 0.771 to 0.778 with the GDINA model generating data; while the Q-matrix recovery rates were more than 0.9 (as high as 1.0) in the case of  $K = 3$  with the DINA model generating data. Even though under the same conditions (10% Q-matrix misspecification rate,  $K = 3$  or  $K = 5$ ), the Q-matrix recovery rate of this study underperformed Chiu’s. It implied that the RSS method might not be suitable for the GDINA-generated data.

The relative index, AIC, showed excellent performance with small samples; therefore, AIC was an appropriate model fit index for small samples. Conversely, BIC was sensitive to Q-matrix misspecification type and sample size; BIC was only suitable for overQ and  $N \geq 200$ . The



absolute indices of the study were sensitive to Q-matrix misspecification type and only displayed excellent performance in the cases of TQ and overQ.

The results showed that in the case of  $K = 3$ , all relative and absolute model fit indices improved model selection with the modified Q-matrix. This implied that Q-matrix modification could improve the performance of model fit indices as few attributes or small domain measured. However, in the case of  $K = 5$ , both modification methods exhibited poor performance. It might be resulted from the complexity of more attribute or the generating GDINA model since the recovery rate of the modified Q-matrix in the case of  $K = 3$  performed better. Meanwhile, we also found similar pattern on the performance of the model fit indices by using the modified Q-matrix. Therefore, these limitations should be taken into consideration in future studies to expand the application of CDMs in practice.

## ORCID

Hueying Tzou  <https://orcid.org/0000-0002-6740-6852>

Ya-Huei Yang  <https://orcid.org/0000-0002-4109-2381>

## 5. REFERENCES

- Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automated Control*, *19*, 716-723.
- DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 979-1030). Amsterdam, Netherlands: Elsevier.
- Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnostic modeling. *Journal of Educational Measurement*, *50*, 123-140.
- Chiu, C.Y. (2013). Statistical refinement of the Q-Matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598-618.
- Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*(2), 225-250.
- de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343-362.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179-199.
- de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*(4), 355-373.
- de la Torre, J., & Chiu, C. Y. (2016). A General Method of Empirical Q-matrix Validation. *Psychometrika*, *81*(2), 253-273.
- Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y. (2016). Evaluation of Model Fit in Cognitive Diagnosis Models. *International Journal of Testing*, *16*(2), 119-141.
- Jiao, H. (2009). Diagnostic classification models: Which one should I use? *Measurement: Interdisciplinary Research & Perspective*, *7*(1), 65-67.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272.
- Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59-81.
- Lei, P.-W., & Li, H. (2016). Choosing Correct Cognitive Diagnostic Models and Q-Matrices. *Applied Psychological Measurement*, *40*(6), 1-12.

- Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model Similarity, Model Selection, and Attribute Classification. *Applied Psychological Measurement, 40*(3), 200-217.
- Ma, W., & de la Torre, J. (2018). GDINA: The generalized DINA model framework. R package version 2.3. Retrived from <https://CRAN.R-project.org/package=GDINA>
- R Core Team (2017). R: A language and environment for statistical computing (Version 3.4.3) [Computing software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org>
- Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification nonparameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78-96.
- Schwarzer, G. (1976). Estimating the dimension of a model. *Annals of Statistics, 6*,461–464.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345-354.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305.

6. APPENDIX

**Table A1.** Selection Rates of the Relative Indices Under Various Simulation Conditions

| K=5 |       | TQ  |       | overQ |       |     |       | underQ |       |       |       | mixQ |       |       |       |
|-----|-------|-----|-------|-------|-------|-----|-------|--------|-------|-------|-------|------|-------|-------|-------|
|     |       |     |       | 5%    |       | 10% |       | 5%     |       | 10%   |       | 5%   |       | 10%   |       |
| N   | M     | AIC | BIC   | AIC   | BIC   | AIC | BIC   | AIC    | BIC   | AIC   | BIC   | AIC  | BIC   | AIC   | BIC   |
| 50  | DINA  | 0   | 0.133 | 0     | 0.333 | 0   | 0.4   | 0      | 0.167 | 0     | 0.167 | 0    | 0.233 | 0     | 0.267 |
|     | DINO  | 0   | 0.867 | 0     | 0.667 | 0   | 0.6   | 0      | 0.833 | 0.067 | 0.833 | 0    | 0.767 | 0.067 | 0.733 |
|     | GDINA | 1   | 0     | 1     | 0     | 1   | 0     | 1      | 0     | 0.933 | 0     | 1    | 0     | 0.933 | 0     |
| 75  | DINA  | 0   | 0.033 | 0     | 0.133 | 0   | 0.3   | 0      | 0.167 | 0     | 0.067 | 0    | 0.2   | 0     | 0.167 |
|     | DINO  | 0   | 0.567 | 0     | 0.767 | 0   | 0.7   | 0      | 0.7   | 0.033 | 0.833 | 0    | 0.733 | 0     | 0.833 |
|     | GDINA | 1   | 0.4   | 1     | 0.1   | 1   | 0     | 1      | 0.133 | 0.967 | 0.1   | 1    | 0.067 | 1     | 0     |
| 100 | DINA  | 0   | 0     | 0     | 0.067 | 0   | 0.133 | 0      | 0     | 0     | 0.033 | 0    | 0.033 | 0     | 0.267 |
|     | DINO  | 0   | 0.2   | 0     | 0.4   | 0   | 0.667 | 0      | 0.333 | 0     | 0.567 | 0    | 0.6   | 0     | 0.733 |
|     | GDINA | 1   | 0.8   | 1     | 0.533 | 1   | 0.2   | 1      | 0.667 | 1     | 0.4   | 1    | 0.367 | 1     | 0     |
| 200 | DINA  | 0   | 0     | 0     | 0     | 0   | 0     | 0      | 0     | 0     | 0     | 0    | 0     | 0     | 0.067 |
|     | DINO  | 0   | 0     | 0     | 0     | 0   | 0     | 0      | 0     | 0     | 0.033 | 0    | 0     | 0     | 0.333 |
|     | GDINA | 1   | 1     | 1     | 1     | 1   | 1     | 1      | 1     | 1     | 0.967 | 1    | 1     | 1     | 0.6   |

Note: M = analytic model

**Table A2.** Selection Rates for Relative Indices with the Modified Q-Matrix

| K=5, RSS |       | overQ |       |       |       | underQ |       |       |       | mixQ  |       |       |       |
|----------|-------|-------|-------|-------|-------|--------|-------|-------|-------|-------|-------|-------|-------|
|          |       | 5%    |       | 10%   |       | 5%     |       | 10%   |       | 5%    |       | 10%   |       |
| N        | M     | AIC   | BIC   | AIC   | BIC   | AIC    | BIC   | AIC   | BIC   | AIC   | BIC   | AIC   | BIC   |
| 50       | DINA  | 0.600 | 0.933 | 0.600 | 0.967 | 0.433  | 0.967 | 0.633 | 0.933 | 0.600 | 0.967 | 0.567 | 1     |
|          | DINO  | 0.033 | 0.033 | 0.033 | 0.033 | 0      | 0.033 | 0.033 | 0.067 | 0     | 0     | 0     | 0     |
|          | GDINA | 0.367 | 0.033 | 0.367 | 0     | 0.567  | 0     | 0.400 | 0.067 | 0.400 | 0.033 | 0.433 | 0     |
| 75       | DINA  | 0.133 | 0.833 | 0.267 | 0.867 | 0.200  | 0.900 | 0.167 | 0.900 | 0.167 | 0.800 | 0.100 | 0.900 |
|          | DINO  | 0     | 0.033 | 0     | 0     | 0      | 0     | 0     | 0     | 0     | 0     | 0     | 0     |
|          | GDINA | 0.867 | 0.133 | 0.733 | 0.133 | 0.800  | 0.100 | 0.833 | 0.100 | 0.833 | 0.200 | 0.900 | 0.100 |
| 100      | DINA  | 0.100 | 0.767 | 0.133 | 0.733 | 0.133  | 0.800 | 0.167 | 0.800 | 0.100 | 0.800 | 0.133 | 0.700 |
|          | DINO  | 0     | 0     | 0     | 0     | 0      | 0     | 0     | 0     | 0     | 0     | 0     | 0     |
|          | GDINA | 0.900 | 0.233 | 0.867 | 0.267 | 0.867  | 0.200 | 0.833 | 0.200 | 0.900 | 0.200 | 0.867 | 0.300 |
| 200      | DINA  | 0.033 | 0.267 | 0.033 | 0.300 | 0.067  | 0.333 | 0     | 0.333 | 0.033 | 0.200 | 0.033 | 0.200 |
|          | DINO  | 0.033 | 0.033 | 0.033 | 0.033 | 0.033  | 0.033 | 0     | 0     | 0.033 | 0.033 | 0     | 0     |
|          | GDINA | 1     | 0.767 | 1     | 0.733 | 0.967  | 0.700 | 1     | 0.667 | 1     | 0.833 | 0.967 | 0.800 |

**Table A3.** Rejection Rates of the Absolute Indices Under Various Simulation Conditions

| K=5 |       | TQ    |       | overQ |       |       |   | underQ |       |       |       | mixQ |   |     |   |
|-----|-------|-------|-------|-------|-------|-------|---|--------|-------|-------|-------|------|---|-----|---|
|     |       |       |       | 5%    |       | 10%   |   | 5%     |       | 10%   |       | 5%   |   | 10% |   |
| N   | M     | r     | l     | r     | l     | r     | l | r      | l     | r     | l     | r    | l | r   | l |
| 50  | DINA  | 0.967 | 0.967 | 1     | 0.967 | 1     | 1 | 0.933  | 1     | 1     | 1     | 1    | 1 | 1   | 1 |
|     | DINO  | 0.800 | 0.900 | 0.967 | 1     | 1     | 1 | 0.867  | 0.967 | 0.967 | 1     | 1    | 1 | 1   | 1 |
|     | GDINA | 0.033 | 0     | 0     | 0     | 0.033 | 0 | 0.667  | 0.800 | 0.967 | 1     | 1    | 1 | 1   | 1 |
| 75  | DINA  | 0.967 | 1     | 1     | 1     | 1     | 1 | 1      | 1     | 1     | 1     | 1    | 1 | 1   | 1 |
|     | DINO  | 1     | 1     | 1     | 1     | 1     | 1 | 1      | 1     | 1     | 1     | 1    | 1 | 1   | 1 |
|     | GDINA | 0     | 0.033 | 0     | 0.033 | 0     | 0 | 0.967  | 0.933 | 1     | 1     | 1    | 1 | 1   | 1 |
| 100 | DINA  | 1     | 1     | 1     | 1     | 1     | 1 | 1      | 1     | 1     | 1     | 1    | 1 | 1   | 1 |
|     | DINO  | 0.967 | 1     | 1     | 1     | 1     | 1 | 1      | 1     | 1     | 1     | 1    | 1 | 1   | 1 |
|     | GDINA | 0     | 0.033 | 0     | 0     | 0     | 0 | 0.933  | 0.967 | 0.967 | 0.967 | 1    | 1 | 1   | 1 |
| 200 | DINA  | 1     | 1     | 1     | 1     | 1     | 1 | 1      | 1     | 1     | 1     | 1    | 1 | 1   | 1 |
|     | DINO  | 1     | 1     | 1     | 1     | 1     | 1 | 1      | 1     | 1     | 1     | 1    | 1 | 1   | 1 |
|     | GDINA | 0     | 0     | 0     | 0     | 0     | 0 | 1      | 1     | 1     | 1     | 1    | 1 | 1   | 1 |

**Table A4.** Rejection Rates for Absolute Indices with the Modified Q-Matrix

| K=5, RSS |       | overQ |       |       |       | underQ |       |       |       | mixQ  |       |       |       |
|----------|-------|-------|-------|-------|-------|--------|-------|-------|-------|-------|-------|-------|-------|
|          |       | 5%    |       | 10%   |       | 5%     |       | 10%   |       | 5%    |       | 10%   |       |
| N        | M     | r     | l     | r     | l     | r      | l     | r     | l     | r     | l     | r     | l     |
| 50       | DINA  | 0.467 | 0.600 | 0.467 | 0.633 | 0.333  | 0.467 | 0.400 | 0.600 | 0.500 | 0.600 | 0.700 | 0.767 |
|          | DINO  | 0.733 | 0.867 | 0.767 | 0.900 | 0.500  | 0.700 | 0.533 | 0.700 | 0.733 | 0.900 | 0.733 | 0.833 |
|          | GDINA | 0.233 | 0.367 | 0.300 | 0.533 | 0.300  | 0.400 | 0.267 | 0.467 | 0.400 | 0.500 | 0.367 | 0.467 |
| 75       | DINA  | 0.533 | 0.633 | 0.733 | 0.767 | 0.400  | 0.667 | 0.367 | 0.533 | 0.633 | 0.800 | 0.667 | 0.800 |
|          | DINO  | 0.733 | 0.800 | 0.833 | 0.933 | 0.600  | 0.700 | 0.600 | 0.767 | 0.800 | 0.800 | 0.833 | 0.900 |
|          | GDINA | 0.433 | 0.500 | 0.600 | 0.633 | 0.367  | 0.500 | 0.333 | 0.500 | 0.467 | 0.600 | 0.567 | 0.733 |
| 100      | DINA  | 0.733 | 0.833 | 0.867 | 0.900 | 0.767  | 0.900 | 0.733 | 0.833 | 0.700 | 0.867 | 0.833 | 0.900 |
|          | DINO  | 0.833 | 0.867 | 0.900 | 0.900 | 0.833  | 0.900 | 0.833 | 0.900 | 0.833 | 0.867 | 0.867 | 0.900 |
|          | GDINA | 0.700 | 0.767 | 0.800 | 0.833 | 0.700  | 0.833 | 0.767 | 0.900 | 0.767 | 0.867 | 0.767 | 0.867 |
| 200      | DINA  | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1     | 1     | 1     | 1     | 1     |
|          | DINO  | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1     | 1     | 1     | 1     | 1     |
|          | GDINA | 1     | 1     | 1     | 1     | 1      | 1     | 1     | 1     | 1     | 1     | 1     | 1     |