IJATE

Volume 7 Issue 3 2020

International Journal of Assessment Tools in Education

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

http://www.ijate.net

e-ISSN: 2148-7456

IJATE

e-ISSN 2148-7456

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

Volume 7 **Issue 3** **2020**

Dr. Izzet KARA
Publisher
International Journal of Assessment Tools in Education
&
Pamukkale University,
Education Faculty,
Department of Mathematic and Science Education,
20070, Denizli, Turkey

Phone : +90 258 296 1036
Fax : +90 258 296 1200
E-mail : ijate.editor@gmail.com

Frequency : 4 issues per year starting from June 2018 (March, June, September, December)
Online ISSN: 2148-7456
Website : http://www.ijate.net/
http://dergipark.org.tr/en/pub/ijate
Design & Graphic: IJATE

Support Contact
Dr. İzzet KARA
Journal Manager & Founding Editor
Phone : +90 258 296 1036
Fax : +90 258 296 1200
E-mail : ikara@pau.edu.tr

International Journal of Assessment Tools in Education (IJATE) is a peer-reviewed and academic online journal.
The scientific and legal responsibility for manuscripts published in our journal belongs to the authors(s).

International Journal of Assessment Tools in Education

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

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

IJATE is indexed in:

- Emerging Sources Citation Index (ESCI),
- Education Resources Information Center (ERIC),
- TR Index (ULAKBIM),
- European Reference Index for the Humanities and Social Sciences (ERIH PLUS),
- Directory of Open Access Journals (DOAJ),
- Index Copernicus International
- SOBIAD,
- JournalTOCs,
- MIAR 2015 (Information Matrix for Analysis of the Journals),
- idealonline,
- CrossRef,
- ResearchBib,

Editors

Editorial Board

TABLE OF CONTENTS

Research Articles

Review Article

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 305–322

https://doi.org/10.21449/ijate.706558

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The Instrument of Teaching Metacognition in Reading Classrooms: The ITMR

Nesrin Ozturk [1,*]

[1]Izmir Democracy University, Department of Educational Sciences, 35140, Izmir, Turkey

ARTICLE HISTORY

Received: Mar. 19, 2020
Revised: July 3, 2020
Accepted: Aug. 08, 2020

KEYWORDS

Metacognition,
Instruction,
Reading,
Assessment,
Thinking

Abstract: Limited influence of metacognition research in mainstream classrooms may stem from a lack of comprehensive pedagogy and/or inconsistent criteria assessing metacognition instruction. For this problem, an instrument designed for metacognition instruction in reading classes was examined. After a systematic and analytic review of broad literature, scale validation procedures were followed. Items that represent observable and measurable teacher-behavior promoting students' metacognition were generated. Next, QUAID examination, expert-, cognitive-, and focus-group interviews were conducted. Data collected from reading teachers via a computer-assisted survey method were analyzed by exploratory factor analysis, Welch's, and Spearman's tests. Findings confirmed that the ITMR had a unidimensional model accounting for 60% of metacognition instruction (α.97). There were no mean differences in metacognition instruction at any elementary grades. The items on the ITMR were also strongly and positively correlated. Thereby, the ITMR can be used to assist and identify classroom metacognition instruction in reading classrooms.

1. INTRODUCTION

Meaning making in reading pertains to actions and interactions of perceptual processes, cognitive skills, and metacognition (Book, Duffy, Roehler, Meloth, & Vavrus, 1985; Doğanay Bilgi & Özmen, 2014; Myers & Paris, 1978). Readers use cognitive strategies for task demands (Doğanay Bilgi & Özmen, 2014; Garner, 1987; Gourgey, 2001) and simultaneously, they employ metacognition for the effectiveness of cognitive resources (Gourgey, 1998, 2001).

Research demonstrated that metacognition can be successfully taught (Ozturk, 2015) and such trainings can help limited and/ or no metacognitive adequacy (Anastasiou & Griva, 2009; Van Keer & Vanderlinde, 2010; Veenman, Van Hout-Wolters, & Afflerbach, 2006). Following metacognition trainings, research found that individuals' vocabulary, reading awareness, skills, comprehension, and performances improve (e.g. Boulware-Gooden, Carreker, Thornhill, & Joshi, 2007; Cross & Paris, 1988; Curwen, Miller, White-Smith, & Calfee, 2010; Muñiz-Swicegood, 1994; Veenman et al., 2006).

CONTACT: Nesrin Ozturk ✉ ozturknesrin@gmail.com Izmir Democracy University, College of Education, Department of Educational Sciences, 35140, Izmir, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

1.1. Problem and Purpose of the Research

Until the early 2000s, metacognition instruction research confirmed beneficiary impacts of various instructional programs, approaches, techniques, and methods. However, as Duffy (2002) emphasized, `research focus must be on thoughtfully adaptive teaching` (p.36). That is, instead of searching for `foolproof` (Duffy, 2002, p.36) practices such as K-W-L, direct explanation, and/or modeling, research must focus on teachers who would possess and improve a mindset of *being metacognitive*. In alignment with Duffy's arguments, Van Keer and Vanderlinde (2010) and recently Baker (2017) highlighted that albeit research in this field, the degree to which mainstream classroom students demonstrate and practice metacognition is not similar to the ones in research. This discrepancy may stem from either an unsatisfied need for the directives to teach metacognition in classrooms (Kerndl & Aberšek, 2012; Veenman et al., 2006) or teachers' instruction that lacks pedagogies of metacognition (e.g. Curwen et al., 2010; Kerndl & Aberšek, 2012; Perry, Hutchinson, & Thauberger, 2008; Thomas & Barksdale-ladd, 2000). Although verbalized slightly different, such a discrepancy put forward the need to lay out a practical understanding of metacognition instruction.

The argument that teachers' instruction lacks practices for metacognition might be strong while there are limitations in extant research assessing metacognition instruction in mainstream classrooms. Such research has not exclusively identified the factors that represent metacognition instruction and has not consistently captured them, yet. That is, research in this realm has operated different indicators for metacognition instruction and has not paid enough attention to the lack of a research-based standardized measure of metacognition instruction as presented in the following section. For these reasons, observable and measureable standardized criteria for metacognition instruction should be developed and then, examined before labelling classroom instruction. Regarding substantial domain-specific nature of metacognition (Papleontiou-louca, 2003; Schraw, 2001; Tishman & Perkins, 1997; Veenman, 2016; Zimmerman, 2000), this study aims to examine the psychometric characteristics of a metacognition instruction instrument for reading classrooms at elementary school level.

1.2. Literature Review

Metacognition pertains to thinking about thinking and it involves metacognitive knowledge, metacognitive strategies, and metacognitive experiences (Flavell, 1979). Metacognitive readers have knowledge about themselves, genres, topics, task demands, and strategies. They can also employ metacognitive strategies; i.e. planning, monitoring, regulating, and evaluating (Bransford, Brown, & Cocking, 2000; Pintrich, Wolters, & Baxter, 2000; Pintrich, 2002) for various task demands. Metacognitive experiences, on the other hand, occur when readers actively engage in higher-order thinking (i.e. strategic reading). That is, strategic reading occurs when individuals think about the text and strategies purposefully, manage task demands and goals actively, and building comprehension, successfully.

To develop the instrument of teaching metacognition in reading classrooms (the ITMR), I studied a broad set of literature (around N=110) including books, research, and conceptual papers on metacognition (N=96), social learning theories (N=5), and gradual release of responsibility framework (N=2). For the limitations of space and focus of the paper, I will shortly declare the categories by which a systematical and analytical review was done in the following.

Initially, I identified and determined how to develop and foster students' metacognition in reading classrooms. For this task, I reviewed (a) metacognition theory, (b) characteristics of metacognitive readers, (c) metacognition assessment of students' competency, (d) meditations on metacognition instruction, (e) empirical research on metacognition instruction, and (f) supplementary instructional practices for metacognition. After reviewing the previous section, I recognized the need to study social learning theories including; (g) social learning theory (Bandura, 1986, 1971), (h) self-regulated learning (Zimmerman, 2000, 2002), and (i) social constructivism (Vygotsky, 1978). Then, as the ultimate goal is to educate self-directed learners, I reviewed (j) gradual release of responsibility framework (Fisher & Frey, 2013; Pearson & Gallagher, 1983). Finally, I also reviewed (k) research studies assessing teachers' metacognition instruction (N=13) to polish the criteria on the ITMR.

By the insights developed reviwing the previous section, I defined a pedagogy of metacognition. A pedagogy of metacognition (PMR) pertains to the instruction for which teachers employ their metacognition, effective instructional practices for teaching metacognition, and metacognition assessment practices by the principles of social learning theories, purposefully. The purpose of a PMR pertains to developing and fostering students' metacognitive autonomy via a gradual release of responsibility trajectory. I also concluded that generic metacognition instruction can be implemented by seven main components. These include (a) fostering students' metacognitive knowledge, (b) scaffolding students' strategic reading, (c) encouraging students' independence with strategic reading, (d) assessing metacognition, (e) adopting goal directedness, (f) integrating language of thinking, and (g) prolonging metacognition instruction (Ozturk, 2017b).

Development of a PMR was compulsory to harmonize the previous theoretical foundations so as to transfer meatcognition instruction into mainstream classrooms. Specfcially, a PMR helped develop behavioral indicators of metacognition instruction. However, such a pedagogy needs cross-checking with the criteria presented by the extant research assessing pedagogical practices of metacognition. Therefore, items on the ITMR can be confirmed for further investigation. In the following, available research that scrutinized specifications with teaching metacognition will be presented.

1.2.1. Literature on metacognition instruction assessment

The purpose of this section was to detect available measurement criteria of metacognition instruction in the literature. By these criteria, the ITMR items developed following a PMR can be confirmed and/or improved, if at all. In this section, available literature assessing metacognition instruction was categorized into two sets; standardized instruments (N=1) and qualitative research (N=10).

1.2.1.1. Standardized measurement instruments

Following an extensive literature review, Wilson and Bai (2010) found that there were no standardized measurement instruments assessing teachers' metacognitive knowledge and pedagogies of metacognition. Therefore, they recruited 105 graduate students who were K-12 teachers majoring in different areas to develop an instrument measuring teachers' understandings, pedagogical knowledge, and beliefs about metacognition. Their confirmatory factor analysis produced a survey of 20 items that can be rated on a 4 point Likert-scale. They found that the items loaded on 4 factors (p >.05) with at least α >.7 for each. This model explains 61% of the variance in teachers' knowledge and pedagogies of metacognition.

This measurement is a domain general instrument and it covered some instructional practices basically divided into two sets; (a) evaluating students' metacognitive processing and (b) teaching students to use metacognitive thinking strategies. The first set included

- teachers' evaluating students' planning the logistics,
- describing the steps and explaining the rationale of each step for a task-completion,
- being aware of their reasoning in completing a task, and
- describing their actions and learning.

The second set of items pertained to teaching students metacognitive thinking strategies by

- providing students with problem-solving activities,
- increasing students' metacognitive knowledge about thinking strategies in relation to specific objectives,
- having students share their thinking,
- facilitating students' discussions on problem solving,
- modelling students thinking processes,
- having students generate questions regarding the content, and
- having students explain the procedures and processes for their answers or task-completion.

Wilson and Bai's (2010) instrument was the first standardized measurement assessing teachers' knowledge and pedagogical understandings of metacognition; however, it posed some limitations. First of all, this survey does not assess what teachers do but what they know (p.286). Moreover, the items are rated on an agreement-scale; therefore, the survey can identify teachers' beliefs about pedagogical understandings of metacognition. Also, this instrument was not specifically designed for reading classrooms. Because behavioral indicators are domain-general, they might be vague for some reading teachers. The survey also includes some hypothetical and/or very specific items (e.g. creating a roller coaster, creating a Venn diagram, and completing an essay on Sherman's March on Atlanta etc.); these may jeopardize the validity of the instrument. Still, although this measure has some limitations and domain-general characteristics, it is used frequently since then.

1.2.1.2. Qualitative studies

The earliest study in qualitative realm was conducted by Kurtz, Schneider, Carr, Borkowski, and Rellinger (1990). In their study, metacognition instruction was assessed by various questions. These questions had participant-teachers (a) make some instructional decisions after reading different scenrios (N=3), (b) react to the scenerios as True or False (N=4), (c) determine instructional techniques or methods (N=2), and rate the statements on a 5-point frequency scale (N=2). However, when these items are examined closely, only 3 of them can identify metacognition instruction. These items include teaching different learning strategies appropriate to different tasks, giving specific instruction for learning strategies, and informing students about benefits of those strategies.

Duffy (1993) also studied teachers' pedagogies of metacognition. Examining lesson-transcripts, interviews with students, and class-observation, Duffy utilized the following criteria for identifying metacognition instruction; teachers' explaining the rationale for learning strategies, modelling strategic reasoning, as well as scaffolding and providing feedback for students' thinking.

Moreover, Zohar (1999) examined teachers' knowledge and practices of metacognition instruction. However, they did not provide any categories or codes for the analysis. Therefore, I coded their findings to identify potential criteria representing metacognition instruction. The findings mostly focused on explicit teaching of thinking skills, holding metacognitive discussions, and modeling thinking as well as reasoning during problem solving.

Thomas and Barksdale-ladd (2000) also did a study with pre-service teachers. They analyzed student-teachers' reflective journals of tutoring to young readers for their instructional approaches. To capture metacognition instruction, they used the following criteria; demonstrating and/or modeling a reading process aloud, children's reading and thinking aloud, and children's doing reflection on what they read.

Bolhuis and Voeten (2001) examined teachers' practices of metacognition instruction. At secondary education level, they did an observation study. To analyze the data, they obtained the following criteria; teachers' explaining learning strategies, questioning students' learning activities and the importance of subject-matter, students' engagement in learning, problem solving and learning strategies, teachers' giving feedback, teachers' coaching students to monitor and evaluate their learning as well as to manage task difficulties, and teachers' informing students about the learning goals and their relevance to out-of-school contexts.

In another observation study, Fisher (2002) studied teachers' instructional practices for metacognition. In this study, Fisher set teachers' modeling thinking skills and demonstrating metacognitive regulation (i.e. showing how to achieve a goal) as the criteria to capture metacognition instruction.

Perry and colleagues (2008) also studied metacognition instruction. Their criterion included teachers' providing students with opportunities to make choices, control challenge, and engage in self-assessment, modeling, using explicit language, and scaffolding learning.

Furthermore, Curwen and colleagues (2010) studied metacognition instruction through classroom-observations and interviews with teachers during a professional development period. They analyzed teachers' explicit comprehension instruction, students' practice and use of comprehension strategies, students' reflections on new ways of thinking, as well as increased student responsibility and ownership of learning. Teachers were also asked to implement some instructional techniques such as activating background knowledge, thinking aloud, using graphic organizers, analyzing text structure, reflecting on writing prompts and content ideas, as well as synthesizing knowledge.

Moreover, Kerndl and Aberšek (2012) examined teachers' competence with metacognition instruction. They did not present data analysis codes, explicitly. Therefore, I coded their findings and found that they mostly focused on teachers' helping students improve metacognitive knowledge and thinking about their cognitive engagements. Also, helping students monitor and evaluate cognitive processes as well as products was paid attention.

Finally, I also examined pre-service teachers' pedagogies of metacognition (Ozturk, 2016). In this study, I used teachers' modeling and/or thinking aloud strategic reading, informed-strategies teaching, scaffolding students' strategic reading, and having students do self-assessment.

Extant studies are crucial to help identify and confirm behavioral indicators of metacognition instruction; however, they pose some limitations. In almost none of these studies, metacognition instruction was sufficiently defined. Moreover, teachers' pedagogies of metacognition were examined divergently and inconsistently. Without a pedagogical framework, each and every study examined various behavioral indicators of metacognition instruction. Those indicators were not defined and contextualized sufficiently, either. Therefore, such methodologies might not help classroom teachers inform and adjust their instruction for metacognition practices, deliberately.

1.2.1.3. Short summary of literature on metacognition instruction assessment

This section aimed to identify extant criteria for metacognition instruction. I realized that while research utilized divergent criteria for metacognition insturtcion, it did not define and/or conceptualize metacognition instruction and its criteria, sufficiently. Still, extent metacognition instruction criteria mostly either aligned with the gradual release of responsibility framework or reflected fundamental principles of social learning theories.

Extant metacognition instruction assessment practices specifically utilized the following; teachers' increasing students' metacognitive knowledge of cognitive strategies and thinking skills, using an explicit language for informed-strategies-teaching, modeling a cognitive endeavor, thinking and/or reasoning during it, demonstrating a reading process and metacognitive regulation, holding metacognitive discussions with students, informing students about learning goals, having students think-aloud their cognitive endeavors and reflect on them, having students practice strategies and thinking skills, having students engage in problem solving, using strategies, controlling challenge and managing task difficulties, coaching students or providing students with scaffolding during cognitive endeavors and feedback for these activities, initiating students' metacognitive discussions, having students do self-assessment, assessing students' metacognitive practices, and having students develop an ownership of learning.

2. METHOD

2.1. Research Design

This study represents a structured-survey research model. The survey was delivered online for (a) people's tendency to give more honest answers (Sue & Ritter, 2012), (b) being less likely to over- and/or over- report behaviors when responding to the statements on one's own (Bradburn, Sudman, & Wansink, 2004), and (c) limiting any aid or influence from the researcher as suggested by (Andres, 2012).

2.2. Validation of the ITMR

To Schwab (1980), scale validation can be complete in three steps; (1) item generation, (2) scale development, and (3) psychometric examination. In the following section, the first two stages will be described; however, the last stage pertains to data analysis. Therefore, it will constitute the results section of this paper.

2.2.1. Item generation

Items for the ITMR were generated after a PMR was developed with a focus on content validity. For this task, teacher-behaviors (i.e. modelling, explaining, and explicitly teaching strategies, teachers' cooperation with students, initiating students' metacognitive discussions, assessing students' metacognitive acts, students' self-assessment, and students' independence with metacognitive endeavors) fostering students' metacognitive behaviors (i.e. planning by task and text evaluation, strategy selection, monitoring, and performance evaluation) were described. Indeed, these behavioral indicators represent the theoretical foundations for what teachers can do to develop metacognitive competencies in students. Then, these behaviors were cross-checked against the previous researches' categories and/or codes of metacognition instruction assessment practices. Following these steps, the initial set of survey items (N=76) was created.

This survey asked respondents to reflect on and rate their firsthand experiences of teaching metacognition in reading classes. All items were positively worded to control the validity of responses and systematic error (Hinkin, 1995). All statements were accompanied with a bipolar rating scale ranging from (0) *not like me* to (100) *exactly like me*. Following these procedures, all items were examined on QUAID (question understanding) to identify unfamiliar words (e.g. explicitly, monitoring, and feedback) that might hinder comprehension.

2.2.2. Scale development

At this stage, the initial items were examined whether and how well they confirm the expectations about the structure and content of the instrument as Hinkin (1995) suggested. For this task, I followed Fowler's (1995) guideline and consulted experts, interviewed with colleagues in the field, and held a discussion session with in-service teachers.

First of all, I held meetings with experts. There were 3 experts whom I consulted for content and construct validity of the survey items. They are distinguished scholars who taught at a Mid-

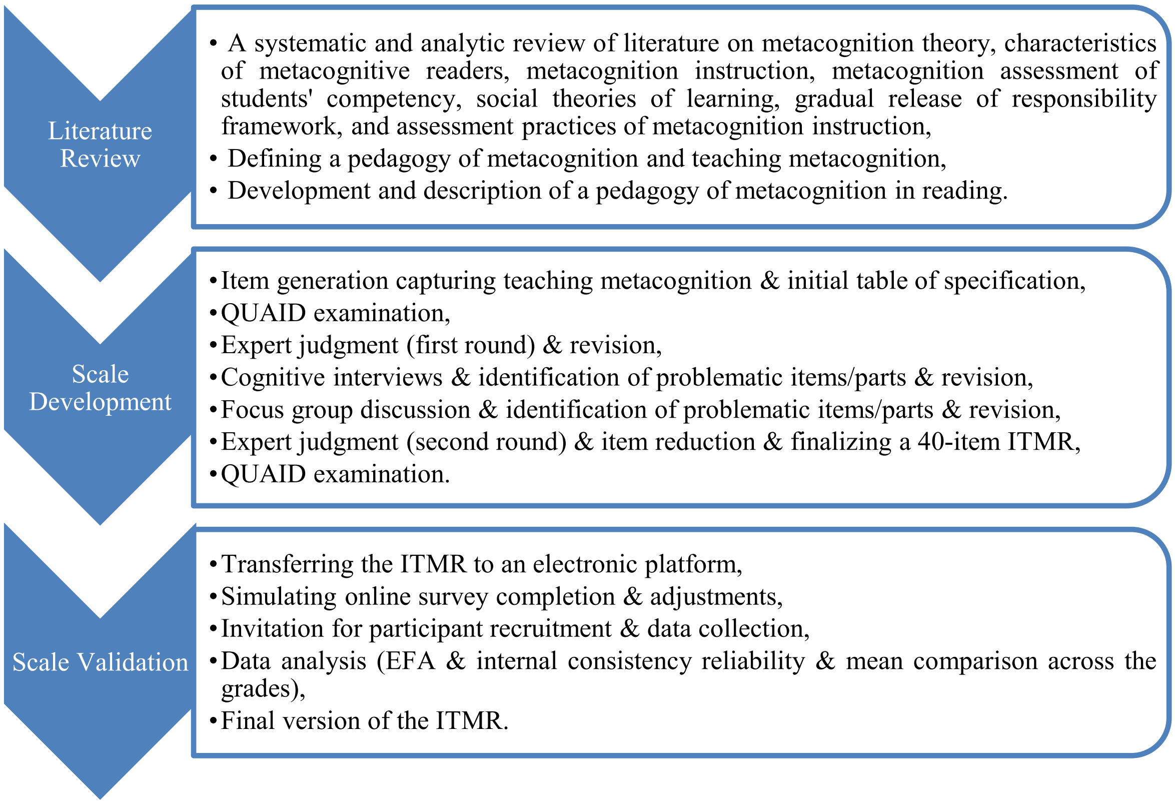
Atlantic public research university in the USA. Each expert had at least 25 years of teaching and research experience in metacognition, strategic processing, strategy teaching, and assessments. Experts were consulted twice for their validity-judgements.

On the first round of expert judgments, I took the initial set and asked whether the survey covers the phenomenon appropriately and reflects its characterization in the domain of reading. Then, I asked them both to respond to the statements and think how potential respondents would comprehend the statements. They were specifically asked whether and what kind of problems the respondents might experience while filling out the survey. Wording of the items were revised based on their feedback. Then, items were presented to cognitive and focus group interview participants.

Following the first round of expert-judgmenet meetings, I held cognitive interviews where colleagues described their thoughts aloud (Fowler, 1995). By cognitive interviews, problems in comprehending the statements, response selection, and appropriateness and relevance of the content can be determined (Fowler, 2009). For these benefits, I held four think-loud interviews with the colleagues (3 females and 1 male). They were familiar with metacognition theory and reading education in the USA. They all held a reading specialist certificate. Their teaching experience ranged from 8 to 13 years. Each interview took around 40 minutes. During each interview, on average 20 statements were studied. The interviewees were specifically asked to paraphrase their understanding of the statements, define the terms, express any confusion or uncertainties while rating the statements, and think about the classroom implementations of the instructional practices. Moreover, participants were also asked how they arrived at choosing a number and how their answers would differ from mainstream classroom teachers. Cognitive interview participants were mostly concerned with the conventions of language. Based on their feedback, I did grammatical revisions. I also took some notes for item-reduction because there were numerous items that sounded very similar.

Along with the cognitive interviews, I also held a focus-group discussion session. Focus group interviews are systematic discussions about the construct under study to identify threads to standardization and to neutralize the complexities that would cause ambiguity (Fowler, 1995). For this study, a relatively homogenous focus-group of eight in-service reading teachers was recruited. At the time of study, they pursued a master's degree in reading education at a Mid-Atlantic public research university. Focus-group participants either taught at elementary (N=5) or middle (N=3) school level. They also had two to eight years of teaching experience. The focus group discussion was conducted during a graduate class. Participants were distributed the initial ITMR and given 30 minutes to study the statements on their own. They were asked to respond to the statements and think whether they would need assistance for clarification. Then, focus-group participants and I discussed the statements for another 30 minutes. I checked QUAID feedback with them and participants reported no problems interpreting the items that included "feedback, explicitly, and monitoring". Therefore, I kept these items for the last version of the ITMR. I also checked the items with focus group for reduction. Following the discussion session, the items to be reduced were identified.

After cognitive interviews and focus-group discussion, I consulted two experts, again. Following the previous procedures, the survey was narrowed down to 40 items representing an intersection of metacognition instruction and students' metacognitive behaviors. Then, these items were transferred to an electronic platform (Qualtrics). Before the survey was delivered to the participants, a few procedures were completed to control any possible factors (i.e. timing, font, and font size) that might impact participants' experiences with the ITMR. The following figure (i.e. Figure 1) presents procedures for the development of the ITMR



Literature Review

- A systematic and analytic review of literature on metacognition theory, characteristics of metacognitive readers, metacognition instruction, metacognition assessment of students' competency, social theories of learning, gradual release of responsibility framework, and assessment practices of metacognition instruction,
- Defining a pedagogy of metacognition and teaching metacognition,
- Development and description of a pedagogy of metacognition in reading.

Scale Development

- Item generation capturing teaching metacognition & initial table of specification,
- QUAID examination,
- Expert judgment (first round) & revision,
- Cognitive interviews & identification of problematic items/parts & revision,
- Focus group discussion & identification of problematic items/parts & revision,
- Expert judgment (second round) & item reduction & finalizing a 40-item ITMR,
- QUAID examination.

Scale Validation

- Transferring the ITMR to an electronic platform,
- Simulating online survey completion & adjustments,
- Invitation for participant recruitment & data collection,
- Data analysis (EFA & internal consistency reliability & mean comparison across the grades),
- Final version of the ITMR.

Figure 1. *Scale validation procedures*

2.3. Data Collection Procedures

Before collecting the data, I made sure that every participant would respond to the same statements, in the same order, and on the same platform to ensure standardization. Following this, I posted a research-invitation to my academic and social networks (e.g. Facebook, LinkedIn, Twitter, ILA, and LRA) to recruit respondents. The invitation included details about the research; purpose, survey completion time, scale type, and participation criteria. To control social desirability, as Bradburn et al. (2004), Fowler, (2009, 1995), Netemeyer, Bearde, and Sharma (2003), as well as Sue and Ritter (2012) suggested, I also assured anonymity and confidentiality of the data. The survey link was active for a month.

2.4. Participants

Target population of this study was specified regarding empirical research practices and theoretical insights. The earliest grade was determined as the first grade regarding Veenman's (2016) and Veenman et al.'s (2006) arguments of that from the age of 8, children can show evidence for metacognitive strategies, efficiently. Considering substantial domain specific manifestations of metacognition, the 5th grade was determined as the upper limit. In addition, regarding Andres's (2012) criteria of grouping unit, geographic boundaries, and time; the sample of this study was narrowed to grade 1 to grade 5 teachers who teach reading in the United States of America during the 2016-2017 academic year. I employed a semi non-probability sampling technique to recruit respondents via online modules because of my limited access to target population. At the end of a month, only 211 of 314 voluntary respondents either satisfied recruitment criteria or provided complete data.

2.4.1. Demographics of the sample

Demographics report respondents' gender, teaching experience, grade, and education level. 211 elementary teachers were dominantly represented by females; there were only nine (4.3%)

males. There were 71 (33.6%) teachers with a bachelor's degree and 140 (66.3%) held a graduate degree. 137 (64.9 %) had a master's and three had a doctoral degree. Respondents taught in various states of the USA; 41 states and D.C. Of these teachers, 34 taught 1st and 5th graders, 35 taught 4th graders, 48 taught 2nd graders, and the rest 60 taught 3rd graders. Teaching experience ranged between a minimum of 1 and a maximum of 40 (years) with a *M*=14.66, *SD*=8.85.

2.4.2. Determination of the sample size

To determine the sample size, I considered recommendations in the literature. To develop a new scale DeVellis (2012), Hinkin (1995), and Nunnally (1978) suggested recruitment of 200, 150, and 300 participants, respectively. Moreover, Bartlett, Kotrlik, and Higgins (2001) emphasized that the ratio of observations to independent variable should not fall below a minimum of 5.

Following data collection, I examined the adequacy of sampling. For this purpose, I conducted an analysis of component saturation regarding de Winter, Dodou, and Wieringa's (2009) and Guadagnoli and Velicer's (1988) recommendations. As de Winter et al., (2009) showed evidence, when the data are well-conditioned with high loadings, small number of factors, and high number of variables; factor analysis can yield reliable results for a sample size. By these criteria and exploratory factor analyses' results, the data set was confirmed adequate.

2.4.3. Post-stratification

Before examining the psychometric properties of the ITMR, I approximated sample's data to the population. For this purpose, I used the most recent data (2011-2012) at the time of this study (Rahman, Fox, Ikoma, & Gray, 2017). Because the sample might diverge from its population, the data were also post-stratified by teachers' education level and Goldring, Gray, and Bitterman's (2013) measures. After this, two iterations of exploratory factor analysis (EFA) were run (original data and weighted data). These two solutions identified the same items constituting the ITMR at elementary school level.

2.5. Data Analysis Procedures

Psychometric examination was the last stage of the scale validation in this study. The data were analyzed for (1) the variation in the items so that it could possibly be explained by fewer factors, (2) possible mean differences in metacognition instruction across elementary school grades, and (3) possible correlations among the instructional practices on the ITMR.

2.5.1. Determination of data's suitability to factor analysis

I examined Kaiser-Meyer-Olkin (KMO), Bartlett's test of Sphericity, and correlation matrix to verify data's suitability to the EFA. I found the factorability adequacy of sampling was satisfactory; the KMO test indicated a value of .953, Bartlett's test of sphericity was significant (χ^2=7105.197, *df*=780, *p* <.05), and all item correlations were significant at *p*< .05.

2.5.2. Determination of the factor numbers

After confirming data's factorability, I conducted a principal component analysis (PCA) to determine the number of the initial factors. For this purpose, I (a) used Eigenvalues (retained factors with eigenvalues ≥1), (b) examined the scree test, and (c) run Monte Carlo PCA for parallel analysis, and (d) considered that a factor is to explain at least 5% of the variance (DeVellis, 2012; Netemeyer et al., 2003). By these criteria, I run a factor analysis. Although I could identify instructional practices for a PMR and although metacognition theory proposes 3 main categories and 3 subsets for metacognitive knowledge and regulation, respectively, I restrained from hypothesizing about the structure of the instruction in mainstream classrooms. That is, it may not be realistic to separate instructional practices from one another in classrooms and these practices might foster different metacognitive components and/or characteristics in different students. Therefore, I run an exploratory factor analysis.

2.5.3. Factor analysis

Following the previous steps of factor extraction, I conducted a principal axis factoring with varimax rotation and determined the most salient items. For this task, I examined the communalities, rotated factor loadings, and considered content validity of the scale. By the criteria that Netemeyer et al. (2003) proposed, I deleted items which load insignificantly (<.45) and items with extremely high loadings (>.90) from the final ITMR.

Moreover, I regarded content validity to retain items. I examined items that contained relevant information for classroom practices of metacognition instruction for its salience. Therefore, I deleted some items (e.g. I have students assess their own text evaluations (e.g. topic, structure, or genre) before reading) although they had communalities ≥ .44. By these procedures, the final ITMR included 24 items and they will be presented in the results section of this study.

2.5.4. Internal consistency reliability

I examined the scale's reliability by internal consistency reliability. The ITMR produced an α.97

2.5.5. Comparison of mean differences

The items were analyzed in groups to identify any grade-level differences. Considering the data's characteristics, I run a non-parametric test (Welch's test and Games-Howell post hoc analysis) to examine the mean differences in metacognition instruction practices across elementary school grades.

3. RESULT

3.1. The ITMR at Elementary School Level

A principal axis factoring with varimax rotation generated a unidimensional (single factor) model that accounted for 60 % of the total variance in metacognition instruction. Item loadings ranged from .865 to .666. The internal consistency reliability was calculated as α=.97. The ITMR had 24 items (Table 1).

Table 1. *The ITMR at elementary school level*

Items	Factor Loadings
I have students demonstrate their independent text evaluations (e.g. topic, structure, or genre) before reading.	.865
I have students demonstrate their independent task evaluations.	.848
I have students assess their own task evaluation.	.835
I have students discuss their text evaluations (e.g. topic, structure, or genre) before reading.	.820
I explicitly teach students how to evaluate their task performance.	.818
I explain why evaluating task performance is important.	.813
I have students assess their own task performance.	.801
I explicitly teach students how to evaluate the task they are given.	.801
I have students discuss their strategies selection for the reading task	.799
I have students assess their own monitoring text understanding during reading.	.798
I have students demonstrate their independent task performance evaluations.	.794

I explain why task evaluation is important for task performance.	.788
I explicitly teach students how to evaluate the text (e.g. topic, structure, or genre) before reading.	.781
I provide feedback on students' strategy selections for the reading task.	.779
I model how I evaluate my task performance.	.778
I help students while they are evaluating the text (e.g. topic, structure, or genre) before reading.	.763
I provide feedback on students' monitoring text understanding during reading.	.758
I provide feedback on students' task performance evaluations.	.749
I have students assess their own strategy selection for the reading task.	.746
I have students discuss their task evaluations.	.693
I help students while they are selecting appropriate reading strategies for the reading task.	.690
I provide feedback on students' text evaluations (e.g. topic, structure, or genre) before reading.	.689
I help students while they are evaluating the task they are given.	.688
I have students demonstrate their independent monitoring text understanding during reading.	.666

3.2. Metacognition Instruction across Elementary School Grades

By a Welch's test, it was confirmed that there were no statistically significant mean differences representing metacognition instruction across any elementary grades, *Fmodel* (4, 88)=1.15 *p*=.34; *Fexplain* (4, 87.88)=.2.25, *p*=.07; *Fexplicitlyteach* (4, 89.6)=942, *p*=.444; *Fscaffoldteach* (4, 90.5)=.702, *p*=.59; *Fscaffolpeer* (4, 90.36)=1.56, *p*=.19; *Fassessteach* (4, 89.6)=1.70, *p*=.156; *Fassesself* (4, 89.97)=.835, *p*=.506, and *Findependet* (4, 90.7)=1.14, *p*=.339.

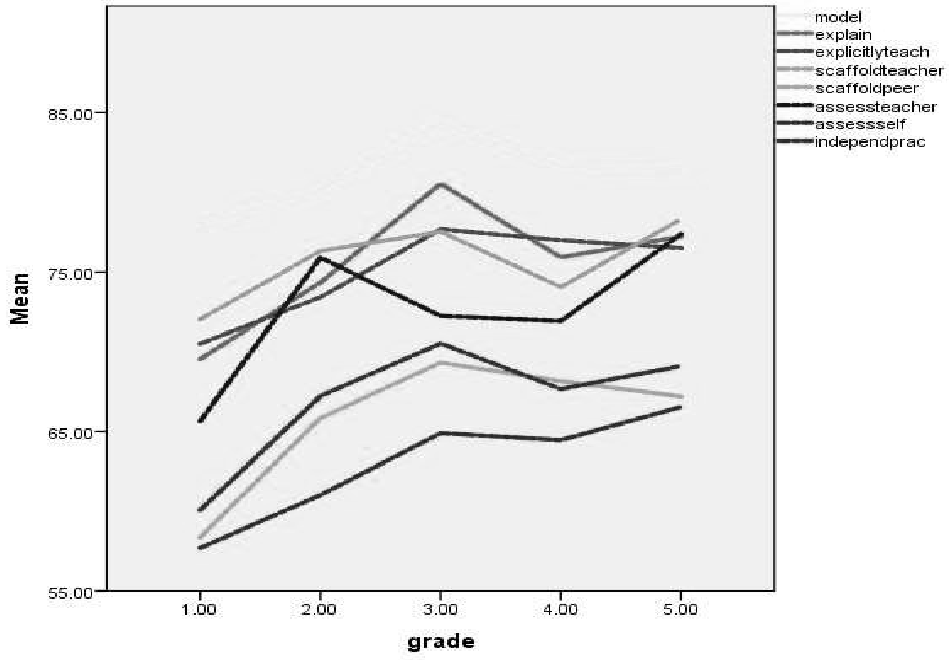


Figure 2. *Metacognition instruction across elementary grades*

3.3. Correlations among ITMR's Items

A series of Spearman's correlation analyses were conducted to examine the relations among the items representing instructional practices on the ITMR. A two-tailed test of significance indicated that all correlation coefficients were statistically significant, strong, and positive, r_s (211) = +.68, $p < .01$.

4. DISCUSSION and CONCLUSION

This study was conducted on the premise of metacognition research's utility for classroom metacognition instruction. Although metacognition research has a long history, the discrepancy between mainstream and research classroom realities regarding students' metacognition competency and proficiency has not been eliminated (Baker, 2017; Carroll, 2008; Curwen et al., 2010; Van Keer & Vanderlinde, 2010). Congruently, teachers' need for practical tools to teach metacognition in classrooms is still not satisfied (Kerndl & Aberšek, 2012). While such problems and needs are still valid, research keeps evaluating classroom metacognition instruction via inconsistent and sometimes, vague criteria. Addressing these urgencies, this study was the first initiative of identifying classroom metacognition instruction in reading classrooms by an instrument; the ITMR. Statistical analyses provided evidence for the ITMR's internal consistency (α.97). The ITMR explained 60% of the total variance in metacognition instruction by a single factor constituting 24 items. The ITMR, currently, may be the only measure of metacognition instruction in the field of reading.

Furthermore, the ITMR can be used across elementary grades. Statistical examination provided evidence for that instructional practices did not show any significant variance at least any two elementary school grades; on the contrary, a similar pattern of metacognition instruction can be observed across all elementary grades. While instructional practices such as modelling, explaining, explicitly teaching, and teacher's scaffolding strategic reading were frequently implemented in mainstream classrooms, students' doing self-assessment was the least frequently implemented practice across all elementary grades.

4.1. Metacognition Instruction: The Literature versus the ITMR

This study identified some discrepancy and congruence between the literature's and the ITMR's criteria representing metacognition instruction and in the following, main findings of this study will be discussed regarding these two sets of criteria.

At elementary school level, the ITMR identified that teaching metacognition was mostly represented by planning (task and text evaluation) and evaluating (task performance). Teachers' presentation behaviors (except for task and performance evaluation) were hardly recognized on the ITMR; however, presentation practices such as teachers' modelling, explaining, and explicitly teaching strategic reading are suggested and highly utilized as the standards of teaching metacognition in literature.

On the ITMR, scaffolding was also identified. In this realm, teachers' scaffolding (via cooperative practices) and peers' scaffolding (by metacognitive discussions) mostly focused on planning reading (task or text evaluation) and regulation of strategies. By identifying reading-phase and/or components, the ITMR helped clarify literature's vague presentation of collaborative practices and scaffolding.

Moreover, literature theoretically proposes comprehension monitoring practices for metacognition instruction. These might include teachers' helping students with comprehension monitoring, students' discussing meaning making with teacher and/or peers, or students' using rubrics (e.g.Collins, Brown, & Holum, 1991; Rosenshine & Meister, 1994). Comprehension monitoring, on the contrary, was the subtlest facet on the ITMR. Approaching the ITMR critically, I have to declare potential influences of educational standards (i.e. Common Core

State Standards) in the context of this study. These standards already require teachers to present and instruct foundational reading skills; therefore, such practices must be common in classrooms.

The most distinctive criteria of classroom metacognition instruction were set by assessment practices. By assessment practices, all stages of strategic reading were identified on the ITMR. That is, students' doing self-assessment of strategic reading and teachers' having students demonstrate task evaluation, text evaluation, comprehension monitoring, and performance evaluation identified on the ITMR. Indeed, these aspects confirm previous arguments (i.e. Lai, 2011; Ozturk, 2017a) of that metacognition is not assessed regularly and traditionally at schools. Therefore, the ITMR's identifying teachers' assessing and then providing feedback on students' strategic reading may not be a coincidence. Moreover, the ITMR's identifying students' doing self-assessment of strategic reading corresponds to the nature of autonomous metacognitive readers as highlighted in the literature (e.g. Afflerbach & Cho, 2009; Afflerbach & Meuwissen, 2005; Veenman et al., 2006).

Lastly, in relation to assessment, students' demonstration of strategic reading organically emerged on the ITMR. For teachers to assess students' strategic reading and for students to reflect on and evaluate strategic reading, students' demonstration of strategic experiences is compulsory. These aspects correspond to the criteria presented in the literature; literature recommends teachers' and students' thinking aloud strategic readings or teachers' evaluating students' reading action plans (e.g. (Baumann, Jones, & Seifert-Kessell, 1993; McKeown & Gentilucci, 2007).

4.2. Metacognition Instruction across Elementary School

This study found that there were no mean differences in metacognition instruction practices across elementary school grades. By so, the ITMR may be applied across all elementary grades. However, the structure of the ITMR reflecting a subtle presence of teacher's presentation of strategic reading and a distinctive proclivity towards assessment practices proposes that classroom teachers might deliver instruction in certain ways.

As seen on Figure 2, teachers' metacognition instruction practices were dived into two distinct sets. On the top, teachers' dominant instructional practices piled up. This set included mostly presentation practices; modeling, explaining, explicitly teaching, and scaffolding students' strategic reading. Therefore, the current classroom trend might be the reason that the ITMR hardly captured such practices.

The least frequently implemented practices pertained to students' agency with strategic reading. This set of practices included encouraging students' demonstration of independent strategic reading, students' scaffolding each other especially via metacognitive discussions, and having students do self-assessment. While these practices were captured by the ITMR, only few researchers including Fisher, (1998, 2007) and Hartman (2001) highlighted utility of metacognitive discussions or dialogic talks for metacognition instruction. Furthermore, as seen on Figure 2, students might not be given enough opportunities to do self-assessment in mainstream classrooms although students gain confidence, mastery, and independence with strategic reading by self-assessment (Afflerbach & Meuwissen, 2005).

Finally, teachers' assessment practices could be blending with or supporting presentation practices as can be interpreted from Figure 2. Although assessing students' strategic reading seems to transpose divergently across the grades, it seems that teachers mostly assessed or utilized the insights while presenting strategic reading or working with students. This is because presentation practices (i.e. modeling, explaining, and explicitly teaching strategic reading) were strongly and positively correlated with teachers' assessment practices.

4.3. Assessment: A Crucial Element of Metacognition Instruction in Reading Classrooms

The discrepancy between the classroom metacognition instruction trends identified in this study and the ITMR's items cannot be ignored regarding assessment practices. Considering demanding educational standards, institutional policies, time pressure, curriculum mandates, high-stake tests, and teachers' expertise with metacognition instruction, it may be that assessment practices were hardly practiced in classrooms. However, the ITMR's criteria highlights the discriminatory importance of assessment (teachers' and students' self-assessment) in developing students' metacognition.

Considering the reciprocal relation between assessment and instruction, teachers' assessing students' metacognition may potentially promote students' metacognition. Teachers can inform and regulate instructional practices for students' metacognition only when they assess students' metacognition competence and needs. After assessing, teachers who are informed about students' current proficiencies with metacognition can implement a need-based and goal-oriented instruction (Ozturk, 2017a). It is after assessing students' metacognition, teachers can decide whether and how to implement metacognition instruction practices practically to address students' extant needs.

Moreover, students' doing self-assessment is the other indispensable pillar of metacognition instruction in classrooms. The purpose of metacognition instruction is to develop students' vicarious control over thinking and their cognitive enterprises (e.g. Papleontiou-louca, 2003; Zimmerman, 2002). Metacognitive readers do self-assessment continuously to test their decisions, behaviors, and impacts of these on and for successful reading experiences. Self-assessment is, in fact, the collection of metacognitive capability (Afflerbach & Meuwissen, 2005); therefore, autonomous readers can direct and control their experiences by doing self-assessment.

5. IMPLICATIONS

5.1. Validity Studies

By the ITMR, this study can initiate a new pathway to study metacognition instruction. First of all, I strongly recommend following a validity-research plan. Messick (1993, 1994) proposed six aspects of validity and this study provided sufficient evidence for content, substantive, and structural validity of the ITMR. Regarding the limitations of this study that stem from data collection procedures (i.e. online) and sampling procedures, The ITMR's use might not be applicable in different settings or its interpretation might be misleading in some settings. For this, I propose future studies to examine (1) the generalizability of the ITMR and to re-run exploratory and/or confirmatory factor analyses before conducting inferential studies. Research should also study (2) the external validity of the ITMR by examining its correlation to other measures. It is also important for future research to examine (3) the consequential validity of the ITMR by especially conducting longitudinal studies. Rather than examining metacognition instruction at a time or for short periods of time, research should study such instructions for sufficient periods to identify the instructional patterns, adequately.

5.2. Educational Implications

While designing this study, I had an altruistic purpose of transferring metacognition litearture to mainstream classrooms, practically. By so, metacognition's beneficial impacts can be observed there. I anticipate this study satisfies teachers' extant needs of metacognition pedagogies and it becomes a supplementary tool. Teachers can adopt the ITMR as a rubric to inform and self-assess their instruction for metacognition.

Moreover, the ITMR can be used to initiate a change in teachers' professional competence. That is, the ITMR can be used to detect the aspects that teachers need scaffolding or improvement. By so, rather than exposing teachers to generic modules of metacognition instruction, needs-based professional development modules at classroom-, school-, or local-levels can be delivered.

Acknowledgements

This study presents partial findings of the doctoral dissertation entitiled *Identifying the Nature of Metacognition Instruction in Reading Classrooms* (Ozturk, 2017b). I would like to express my sincere gratitude and deepest appreciation to my professors, colleagues, friends, and participants who made this study possible.

Declaration of Conflicting Interests and Ethics

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

ORCID

Nesrin Ozturk https://orcid.org/0000-0002-7334-8476

6. REFERENCES

Afflerbach, P., & Cho, B.-Y. (2009). Identifying and describing constructively responsive comprehension strategies in new and traditional forms of reading. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 69–90). New York, NY: Routledge.

Afflerbach, P., & Meuwissen, K. (2005). Teaching and learning self-assessment strategies in middle school. In S. E. Israel, C. Collins Block, K. L. Bauserman, & K. Kinnucan-Welsch (Eds.), *Metacognition in literacy learning: Theory, assessment, instruction, and professional development* (pp. 141–164). Mahwah, NJ: Erlbaum.

Anastasiou, D., & Griva, E. (2009). Awareness of reading strategy use and reading comprehension among poor and good readers. *Elementary Education Online*, *8*(2), 283–297.

Andres, L. (2012). *Designing and doing survey research*. London, England: SAGE.

Baker, L. (2017). The development of metacognitive knowledge and control of comprehension: Contributors and consequences. In K. Mokhtari (Ed.), *Improving reading comprehension through metacognitive reading strategies instruction* (pp. 1–31). Lanham, MD: Rowman & Littlefield.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, Albert. (1971). *Social learning theory*. Morristown, NJ: General Learning.

Bartlett, J. E., Kotrlik, J. W. K. J. W., & Higgins, C. (2001). Organizational research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, *19*(1), 43–50.

Baumann, J. F., Jones, L. A., & Seifert-Kessell, N. (1993). Using enhance comprehensi monitoring The authors think chil ilouds dren ' s ion al program for teaching students think aloud during reading as a means. *The Reading Teacher*, *47*(3), 184–193.

Bolhuis, S., & Voeten, M. J. . (2001). Toward self-directed learning in secondary schools: What do teachers do? *Teaching and Teacher Education*, *17*(7), 837–855.

Book, C., Duffy, G. G., Roehler, L. R., Meloth, M. S., & Vavrus, L. G. (1985). A study of the relationship between teacher explanation and student metacognitive awareness during reading instruction. *Communication Education*, *34*, 29–36.

Boulware-Gooden, R., Carreker, S., Thornhill, A., & Joshi, R. M. (2007). Instruction of metacognitive strategies enhances reading comprehension and vocabulary achievement of third-grade students. *The Reading Teacher*, *61*(1), 70–77.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitve guide to questionnaire design for market research, political polls, and social and health questionnaires* (Revised). San Francisco, CA: Jossey-Bass.

Bransford, J., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school* (Expanded). Washington DC: National Academy.

Carroll, M. (2008). Metacognition in the classroom. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 411–427). New York: Psychology Press.

Collins, A., Brown, J. S., & Holum, A. (1991). Cognitive apprenticeship: Making thinking visible. *American Educator*, *15*(3), 6–11.

Cross, D. R., & Paris, S. G. (1988). Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology*, *80*(2), 131–142. https://doi.org/10.1037/0022-0663.80.2.131

Curwen, M. S., Miller, R. G., White-Smith, K. A., & Calfee, R. C. (2010). Increasing teachers' metacognition develops students' higher learning during content area literacy instruction: Findings from the read-write cycle project. *Issues in Teacher Education*, *19*(2), 127–151.

de Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*(November), 147–181.

DeVellis, R. F. (2012). *Scale Development: Theory and applications* (3rd ed.). Thousand Oaks, CA: SAGE.

Doğanay Bilgi, A., & Özmen, E. R. (2014). The impact of modified multi-component cognitive strategy instruction in the acquisition of metacognitive strategy knowledge in the text comprehension process of students with mental retardation. *Educational Sciences: Theory & Practice*, *14*(2), 707–714.

Duffy, G. G. (1993). Rethinking strategy instruction: Four teachers' development and low achievers' understandings. *Elementary School Journal*, *93*(3), 231.

Duffy, G. G. (2002). The case for direct explanation of strategies. In C. C. Block & M. Pressley (Eds.), *Comprehension instruction: Research-based best practices* (pp. 28–41). New York: Guilford.

Fisher, D., & Frey, N. (2013). *Better learning through structured teaching: A framework for the gradual release of responsibility* (2nd ed.). Alexandria, VA: ASCD.

Fisher, Robert. (1998). Thinking about thinking: Developing metacognition in children. *Early Child Development and Care*, *141*(1), 1–15.

Fisher, Robert. (2007). Dialogic teaching: Developing thinking and metacognition through philosophical discussion. *Early Child Development and Care*, *177*(6–7), 615–631.

Fisher, Ros. (2002). Shared thinking: Metacognitive modelling in the literacy hour. *Reading*, *36*(2), 63–67.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911.

Fowler, F. J. (2009). *Survey research methods* (4th ed.). Thousand Oaks, CA: Sage.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: SAGE.

Garner, R. (1987). *Metacognition and reading comprehension*. Norwood, NJ: Ablex.

Goldring, R., Gray, L., & Bitterman, A. (2013). *Characteristics of Public and Private Elementary and Secondary School Teachers in the United States: Results From the 2011–12 Schools and Staffing Survey (NCES 2013-2014).* Washington DC: NCES, IES, U.S. Department of Education. Retrieved from https://nces.ed.gov/pubs2013/2013314.pdf

Gourgey, A. F. (1998). Metacognition in basic skills instruction. *Instructional Science*, *26*, 81–96.

Gourgey, A. F. (2001). Metacognition in basic skills instruction. In H. J. Hartman (Ed.), *Metacognition in learning and instruction: Theory, research, and practice* (pp. 17–32). Boston: Kluwer.

Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265–275.

Hartman, H. J. (2001). Developing students' meatcognitive knowledge and skills. In H. J. Hartman (Ed.), *Metacognition in learning and instruction: Theory, research, and practice* (pp. 33–68). Boston: Kluwer.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, *21*(5), 967–988.

Kerndl & Aberšek, M. K. (2012). Teachers' competence for developing reader's reception metacognition. *Problems of Education in the 21st Century*, *46*(1979), 52–61.

Kurtz, B. E., Schneider, W., Carr, M., Borkowski, J. G., & Rellinger, E. (1990). Strategy instruction and attributional beliefs in West Germany and the United States: Do teachers foster metacognitive development? *Contemporary Educational Psychology*, *15*(3), 268–283. https://doi.org/http://dx.doi.org/10.1016/0361-476X(90)90024-U

Lai, E. R. (2011). *Metacognition: A Literature review (Research report)*. New York, NY:Pearson. Retrieved from http://www.datec.org.uk/CHAT/chatmeta1.htm

McKeown, R. G., & Gentilucci, J. L. (2007). Think-aloud strategy: Metacognitive development and monitoring comprehension in the middle school second-language classroom. *Journal of Adolescent & Adult Literacy*, *51*(2), 136–147.

Messick, S. (1993). *Foundations of validity: Meaning and consequences in psychological assessment* (RR-93-51). Princeton, New Jersey: Educational Testing Service.

Messick, S. (1994). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning* (RR-94-45). Princeton, New Jersey: Educational Testing Service.

Muñiz-Swicegood, M. (1994). The effects of metacognitive reading strategy training on the reading performance and student reading analysis strategies of third grade bilingual students. *Bilingual Research Journal*, *18*, 83-97. https://doi.org/10.1080/15235882.1994.10162659

Myers, M., & Paris, S. G. (1978). Children's metacognitive knowledge about reading. *Journal of Educational Psychology*, *70*(5), 680–690.

Netemeyer, R. G., Bearde, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: SAGE.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Ozturk, N. (2015). A short review of research on metacognition training. *Journal of Educational and Instructional Studies in the World*, *5(3)*, 50–62.

Ozturk, N. (2016). An analysis of pre-service elementary teachers' understanding of metacognition and pedagogies of metacognition. *Journal of Teacher Education and Educators*, *5*(1), 47–68.

Ozturk, N. (2017a). Assessing metacognition: Theory and practices. *International Journal of Assessment Tools in Education*, *4*(2), 134–148.

Ozturk, N. (2017b). *Identifying the nature of metacognition instruction in reading classrooms (Unpublished doctoral dissertation)*. University of Maryland, College Park, Maryland.

Papleontiou-louca, E. (2003). The concept and instruction of metacognition. *Teacher Development*, *7*(1), 9–30. https://doi.org/10.1080/13664530300200184

Pearson, P. D., & Gallagher, G. (1983). The gradual release of responsibility model of instruction. *Contemporary Educational Psychology*, *8*, 112–123.

Perry, N. E., Hutchinson, L., & Thauberger, C. (2008). Talking about teaching self-regulated learning: Scaffolding student teachers' development and use of practices that promote self-regulated learning. *International Journal of Educational Research*, *47*, 97–108.

Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). Assessing metacognition and self-regulated learning. In Gregory Schraw & J. C. Impara (Eds.), *Assessing metacognition and self-regulated learning* (pp. 43–97). Lincoln, NE: Buros Institute of Mental Measurements.

Pintrich, Paul R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory Into Practice*, *41*(4), 219–225.

Rahman, T., Fox, M. A., Ikoma, S., & Gray, L. (2017). *Certification Status and Experience of U.S. Public School Teachers: Variations Across Student Subgroups (NCES 2017-056)*. Washington, DC: U.S. Government Printing Office. Retrieved from https://nces.ed.gov/pubs2017/2017056.pdf

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research*, *64*(4), 479–530.

Schraw, G. (2001). Promoting general metacognitive awareness. In H. J. Hartman (Ed.), *Metacognition in learning and instruction: Theory, research, and practice* (pp. 3–16). Boston, MA: Kluwer.

Schwab, D. P. (1980). Construct validty in organization behavior. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol 2, pp. 3–43). Greenwich: JAI.

Sue, V. M., & Ritter, L. A. (2012). *Conducting online survey* (2nd ed.). Thousand Oaks: SAGE.

Thomas, K. F., & Barksdale-ladd, M. A. (2000). Metacognitive processes: Teaching strategies in literacy education courses. *Reading Psychology*, *21*, 67–84.

Tishman, S., & Perkins, D. (1997). The language of thinking. *The Phi Delta Kappan*, *78*(5), 368–374.

Van Keer, H., & Vanderlinde, R. (2010). The impact of cross-age peer tutoring on third and sixth graders' reading strategy awareness, reading strategy use, and reading comprehension. *Middle Grades Research Journal*, *5*(1), 33–45.

Veenman, M. V. J. (2016). Metacognition. In P. Afflerbach (Ed.), *Handbook of individual differences in reading* (pp. 26–40). Routledge.

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*(1), 3–14. https://doi.org/10.1007/s11409-006-6893-0

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wilson, N. S., & Bai, H. (2010). The relationships and impact of teachers' metacognitive knowledge and pedagogical understandings of metacognition. *Metacognition and Learning*, *5*(3), 269–288. https://doi.org/10.1007/s11409-010-9062-4

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). San Diego, CA: Academic.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An Overview. *Theory Into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102

Zohar, A. (1999). Teachers' metacognitive knowledge and the instruction of higher order thinking. *Teaching and Teacher Education*, *15*(4), 413–429.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 323–342
https://doi.org/10.21449/ijate.707199

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

Development of a Multidimensional Computerized Adaptive Test based on the Bifactor Model

Murat Dogan Sahin[1,*], **Selahattin Gelbal**[2]

[1]Anadolu University, Department of Educational Sciences, Eskişehir, Turkey
[2]Hacettepe University, Department of Educational Sciences, Ankara, Turkey

ARTICLE HISTORY

Received: Mar. 19, 2020
Revised: July 3, 2020
Accepted: Aug. 15, 2020

KEYWORDS

Multidimensional Computerized Adaptive Testing,
Real-time Application,
Bifactor Model,
Hybrid Simulation

Abstract: The purpose of this study was to conduct a real-time multidimensional computerized adaptive test (MCAT) using data from a previous paper-pencil test (PPT) regarding the grammar and vocabulary dimensions of an end-of-term proficiency exam conducted on students in a preparatory class at a university. An item pool was established through four separate 50-item sets applied in four different semesters. The fit between unidimensional, multi-unidimensional and bifactor IRT models was compared during item calibration, with the bifactor model providing the best fit for all data sets. This was followed by a hybrid simulation for 36 conditions obtained using six item selection methods, two ability estimation methods and three termination rules. The statistics and graphs obtained indicate D-rule item selection, maximum a posteriori (MAP) ability estimation and standard error termination rule as the best algorithm for the real-time MCAT application. With the minimum number of items to be administered determined as 10, the real-time application conducted on 99 examinees yielded an average number of items of 13.4. The PPT format proficiency exam consists of 50 items, leading to the conclusion that the examinees participating in the real-time MCAT are administered an average of 74.4% fewer items than the PPT. Additionally, 86 of the examinees answered between 10-13 items. The item pool use rate is 30%. Lastly, the correlation between the PPT scores and general trait scores of 32 examinees was calculated as .77.

1. INTRODUCTION

The development of applications based on rapid and constant data flow has added momentum to studies on rapidly obtaining measurements from individuals and minimizing error levels in these measurements. To this end, it may be stated that measurement practices based on advanced technologies have gained importance from a psychometric perspective.

When measuring a trait of an individual, standard tests are commonly utilized. Due to the ease of application and to ensure understanding among individuals not versed in psychometry literature, Classical Test Theory (CTT) is frequently used for the development of these tests (Jabrayilov, Emons & Sijtsma, 2016). However, while CTT provides ease in practical application and evaluation, it carries many limitations from a psychometric perspective. It may

CONTACT: Murat Doğan ŞAHİN ✉ muratdogansahin@gmail.com 🖃 Anadolu University, College of Education, Department of Educational Sciences- Eskişehir, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

be stated that Item Response Theory (IRT) addresses the theoretical limitations of CTT (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000). IRT posits that the estimated ability parameters are independent from the items administered to individuals. Given that test scores are equalized, this feature allows for the comparison of individuals' abilities independent from the item group (Kelecioğlu, 2001).

IRT states that just as item and ability parameters are independent from the group, standard error can be obtained for the estimated ability level of each separate individual. In addition to that characteristic, IRT also posits unidimensionality and the local independence that emerges as a result of this must be ensured to conduct scaling (van der Linden, 2016). Despite the fact that IRT is based on the assumption of unidimensionality, accepting that scales measure a single dominant latent variable contradicts the multidimensional nature of psychological constructs in practice (Reise, Morizot & Hays, 2007). Therefore, through the expansion of unidimensional IRT, multidimensional IRT emerged (Bock & Aitkin, 1981).

Due to the sophisticated mathematical foundation required by IRT, the development of the theory was stagnant until the end of the 1960's. A dominance of scientific work on IRT was observed in the 1970's (Hambeleton & Swaminathan, 1985). From this day onward, in addition to studies contributing to the theoretical development of IRT, studies were conducted comparing the ability estimations based on either CTT or IRT, obtained from the findings of tests applied to individuals. These studies indicate high correlation between IRT and CTT ability estimations for both unidimensional and multidimensional models (Gelbal, 1994; Fan, 1998; Progar & Sočan, 2008; Çelen & Aybek, 2013; Ferrando & Chico, 2007, Lawson, 1991; Ndalichako & Rogers, 1997; Akyıldız & Şahin, 2017). This situation raises the question of necessity regarding the scaling of PPT in accordance with IRT due to the complex mathematical foundations it requires. Some psychometrists posit that the purpose of IRT's existence lies in Computerized Adaptive Testing (CAT) applications (Weiss, 1985; Wainer et al., 2000; Ware et al., 2003).

Using a precalibrated item pool, CAT is an application that is based on making a provisional ability estimation for the examinee, selecting and applying the item from the pool most appropriate for the provisional ability estimation, and concluding the test in accordance with a predetermined rule (Frey, 2009; Thompson & Weiss, 2011; Bulut & Kan, 2012). A diagram of the realization of a CAT application is presented in Figure 1.

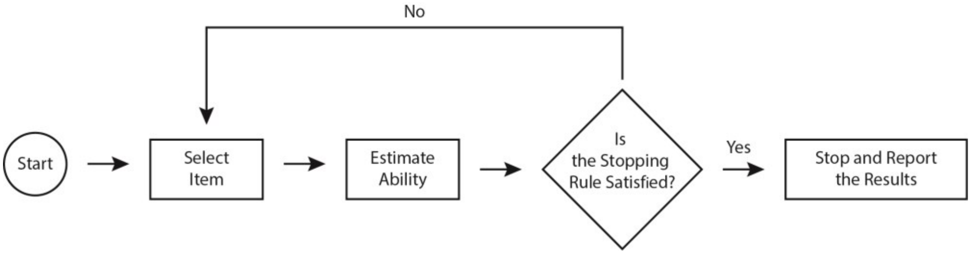


Figure 1. *CAT Applications Flow Chart*

In CAT applications; as the individuals only respond to items appropriate for their provisional ability levels, a measurement accuracy identical to a standard test that applies to the whole group is obtained through much fewer items being applied (Segall, 2005; Weiss, 2011). The ability to present individuals with items appropriate for their level in CAT applications is based on the fact that the ability level of an individual rests on the same scale as item difficulty within the scope of IRT (Reckase, 2009). Studies indicate that CAT applications provide the same measurement accuracy as PPT with 50% fewer items on average (Segall, 1996; Luecht, 1996; Eggen, 2007; Weiss & Gibbons, 2007; Gibbons et al., 2008; Weiss, 1985, 2011; Kalender & Berberoğlu, 2016).

The majority of studies on CAT applications were developed based on unidimensional IRT. However, developments in computer technologies have been increasing the interest in multidimensional CAT studies (Reckase, 2009).

Following studies in the field aiming to increase the measurement accuracy of multidimensional CAT (MCAT) compared to unidimensional CATs (e.g. Segall, 1996; Luecht, 1996), research aiming to increase the efficiency of MCAT applications grew in prominence (e.g. Veldkamp & van der Linden, 2002; Wang & Chen, 2004; Mulder & van der Linden, 2009). In the past decade, multiple studies have been conducted on developing methods regarding MCAT applications such as item selection, test termination, content balancing, etc. (Choi, Grady & Dodd, 2010; Yao, 2012, 2013, 2014; Wang, Chang & Boughton, 2012; Yao, Pommerich & Segall, 2014; Su, 2016; Lin & Chang, 2019). These studies are noted to mainly focus on within-item or between-item dimensionality. Beyond these studies, there appears to be limited research in which MCAT studies execute general trait estimation that take into account the common source of variance underlying the dimensions (sub factors) that establish the items or structure without disregarding multidimensionality (Weiss & Gibbons, 2007; Seo, 2012; Huang, Chen & Wang, 2012; Seo & Weiss, 2015; Zheng, Chang & Chang, 2013).

The purpose of this study is to portray the applicability of a PPT used to measure the grammar and vocabulary dimensions of the English proficiency of university students, following a preparatory class as an MCAT. The study consists of three main sections, in the first of which items from the proficiency exam conducted in various years as a PPT are calibrated to create an item pool. The second section consists of a hybrid simulation based on the sparse data matrix completed as a result of the missing responses created from the estimated ability levels of individuals, and the best condition for a real-time MCAT application is portrayed. The final section consists of the real-time MCAT application conducted in accordance with the algorithm based on simulation results.

In MCAT applications, multidimensional IRT models that fundamentally rely on within-item or between-item dimensionality models are used. The between-item dimensionality model (also known as multi-unidimensional model) accepts that each item measures only one dimension; however this situation is unrealistic when the nature of psychological structures are considered. The within-item model, however, assigns weight to all dimensions. In these models though, the definability of dimensions is problematic (Li & Schafer, 2005). The bifactor model used in this study provides a solution for related structures foreseen to have a general factor/ability (general trait) (Gustafson & Balke, 1993). When evaluating multidimensional constructs in order to provide the domain score, the bifactor model is considered to be highly relevant (Nieto, Abad & Olea, 2018). As such, it may be stated that this study suits the nature of English proficiency in that it will provide a general trait estimation without disregarding multidimensionality.

Thompson and Weiss (2011) state that the most important advantage of CAT applications is that they place the ability level of an individual on the same scale as item difficulty, ensuring the selection of items appropriate for the ability level of the individual being measured by the test. This ensures that individuals are only required to answer items suitable to their ability levels, resulting in a test concluded with much fewer items than they would have answered with a traditional PPT. This adaptation of the test to the individual negates the need for individuals to respond to items above or below their ability levels thereby minimizing standard error of measurement and increasing the measurement accuracy. In other words, CAT applications achieve the same measurement accuracy as traditional tests with much fewer items (Gibbons et al., 2008; Weiss, 2011). Segall (2005) states that the increase in measurement efficiency of CAT applications depends on the measurement accuracy and the length of the test, while Weiss (2011) indicates that an increase in measurement accuracy is directly related to the reduction in the number of items administered.

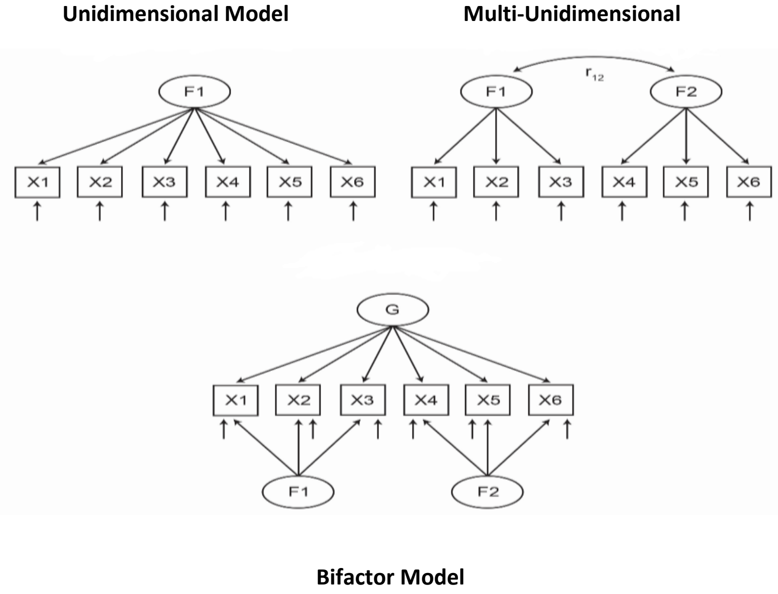
The fundamental components of a CAT application are; a calibrated item pool, starting rule, item selection method, ability estimation method and termination (stopping) rule (Weiss & Kingsbury, 1984; Thompson & Weiss, 2011). Beyond these components, item exposure for the effective use of the item pool, and content balancing methods for a balanced representation of item scope may be used. However, in situations where the item pool is small, the use of item exposure dramatically increases the number of items administered to due to the limited number of items reducing the number of items equivalent to each other in terms of information function (Huebner et al., 2016). Therefore, this study does not use the item exposure method. Due to the fact that the bifactor model provides equal distribution among specific factors and their related items by default, there was no need to use any content balancing method.

2. METHOD

This study may be divided into three segments, namely calibrating the item pool, hybrid simulation, and real time MCAT application.

2.1. Item Pool Calibration

The item pool consists of 200 questions developed to measure grammar and vocabulary skills, applied at the end of a university preparatory class. Each 50 of these 200 questions were applied between 2014-2016, at the end of four different semesters. The 50 item sets were conducted on 415, 692, 798, and 1153 students in that order. During item preparation, English Language Teaching experts who have an experience of instruction and question preparation at a proficiency level contributed to the preparation, and items were prepared in accordance with the Global Scale of English (GSE) developed by Pearson.


Unidimensional Model

Multi-Unidimensional

Bifactor Model

Figure 2. *Figures of the IRT Models Used in this Study*

Within the scope of this research, a multidimensional item response theory (MIRT) package (Chalmers, 2012) defined in R was used to calibrate four data sets in accordance with unidimensional, multi-unidimensional (between-item dimensionality), and bifactor models (see Figure 2). In each of these three models, 2PL was used. For each 50 item sets, a likelihood ratio chi-square statistic was used to determine whether the bifactor model improved fit over unidimensional and multi-unidimensional alternatives. It was concluded that the most appropriate approach was the bifactor model for each of the four item sets. As a result of the applications conducted to portray the invariance of the item and ability parameters, it was observed that the the correlations between the item parameters for the lower and upper groups, and the correlations between the ability estimations determined from randomly assigning two groups of item sets were statistically significant.

2.2. Hybrid Simulation

Following the establishment of the item pool, hybrid simulation was conducted. Post-hoc simulation applications based on data obtained from the PPT application of items are used to decide the different initiation, provisional estimate of ability level, and termination rules to be used in the algorithm for the application (Weiss, 2004). During post-hoc simulations; the responses examinees provide to the items in the PPT format that establish the CAT pool are accepted as the responses they provide for the same item in the CAT application (Nydick & Weiss, 2009). Therefore, post-hoc simulations are also called "real data" simulations (Thompson & Weiss, 2011). However, the ability of a post-hoc simulation to correctly estimate a CAT output depends on all items being answered by all examinees (Weiss & Gibbons, 2007; Gibbons et al., 2008). Additionally, a complete response matrix in which examinees respond to all items cannot be obtained if item sets are applied to different groups. In such instances, completing the sparse response matrix through hybrid simulation is appropriate. Hybrid simulations use monte carlo and post-hoc simulations together to seek an answer to this question: "what would happen if all the examinees responded to all the items in the item pool?". This approach means that this question set can be tested for CAT function without the need for all items to be administered to all examinees, despite there being examinees in different groups whom have not answered some of the items in the pool.

Since the item pool in this study consists of four separate item sets applied to different groups, first, examinees' ability levels were estimated based on the 50 items they responded to, then their missing responses for the other three item sets in the sparse response matrix were generated based on their ability levels and the parameters of these items. The real and generated responses were then combined to create a 3058*200 response matrix. In turn, this matrix was used to calculate the correlation, bias, RMSD, and standard error among the θ values estimated from the PPT and hybrid simulation for 36 different conditions (see Table 1). The average number of items administered was also reported, as it is an important indicator of measurement accuracy in variable length applications. In the termination rules depending on variable test length, the minimum number of items to be administered based on the opinions of experts regarding content validity was determined as 10, while the maximum number of items in the instance that termination conditions could not be established was determined as 60. mirtCAT (Chalmers, 2016) was used for hybrid simulation applications. The initiation rule mandated by this package. It was the determination of a fixed item, therefore an item from the item pool with medium difficulty and high discrimination levels was chosen as a test initiation rule for all applications.

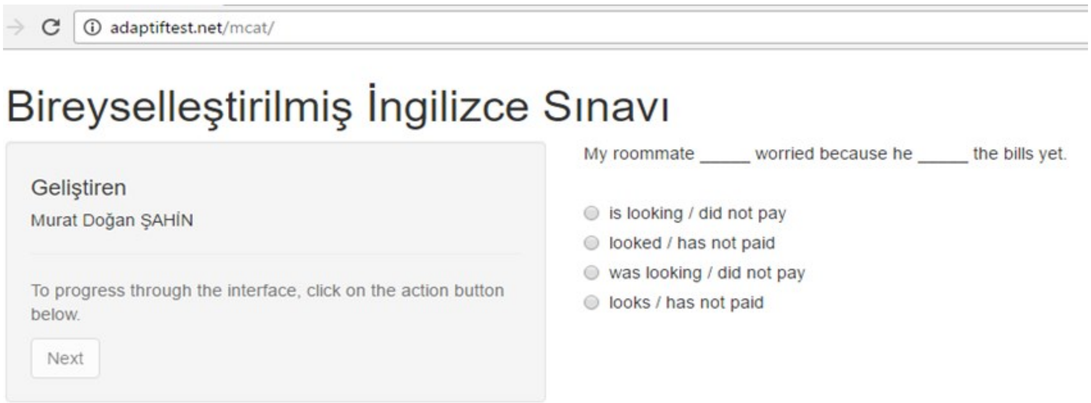
Table 1. *CAT Components Establishing 36 Conditions in the Simulation*

CAT Components	Method	Number of Conditions
Ability Estimation	EAP (expected a posteriori) ve MAP (maximum a posteriori)	2
Item Selaction	D-rule (the determinant rule), KL (the Kullback-Leibler divergence criteria), W-rule (weighted composite rule), weighted* W-rule, T-rule (trace of the information)and weighted* T-rule	6
Termination Rule	Standard error (.40), θ convergence ($\Delta\theta < .05$) ve fixed number of items (k=20)	3

* The weighting was determined to be for the general trait (.8, .1, .1).

2.3. Real-Time MCAT Application

In the final stage of the study, the best condition determined based on the hybrid simulation was the algorithm of the real-time MCAT application. The real-time MCAT application was conducted at the end of the preparatory class with 99 students (47 female, 52 male; age=19.3), taking advantage of the mirtCAT (Chalmers, 2016) package defined in R. For a graphical user interface (GUI), the shiny (Chang, 2019) package defined in R was used, and the researcher used their personal server during the application. An example for the interface encountered by the responder during the application is portrayed in Figure 3.



adaptiftest.net/mcat/

Bireyselleştirilmiş İngilizce Sınavı

Geliştiren

Murat Doğan ŞAHİN

To progress through the interface, click on the action button below.

My roommate _____ worried because he _____ the bills yet.

○ is looking / did not pay
○ looked / has not paid
○ was looking / did not pay
○ looks / has not paid

Figure 3. *GUI image of the real-time application*

It is notable that the "next" button is not active in the image above. This is due to the fact that despite not being encountered in the local application, the application enters an error state if the next button is clicked without a response to the item in the online application. As the application does not continue where it left off when this error is encountered, and a new examination application is not allowed without refreshing the server, a "javascript" applet was written to activate the "next" button when the item was responded to. The application lasted between 9-13 minutes for each student. Only one student who responded to 45 items took an 18-minute duration.

The results reflected in the database following the application show the final θ for each examinee, the standard error values for these θs, the responses for each item, the status of these responses (1-0), the ID's of the items in the database, θ and standard error histories, and lastly the time spent to respond for each item. Additionally, the correlations between the total PPT scores and the general trait scores from the real-time MCAT application of 32 students were calculated, and statistics regarding the use state of the item pool were shared.

Based on all of these practices, the research problems that emerged were as follows:

1. For the 36 different conditions within the scope of the research, taking into consideration error statistics and average number of items administered, which condition is the best for the real-time MCAT application?
2. What are the real-time MCAT application results regarding number of items administered, use rates of the item pool, and examinees' θ estimations obtained from PPT and MCAT?

3. RESULT

3.1. Hybrid Simulation Results

The results of the 36 conditions determined for the simulation were reported based on the termination rules. A study of the results obtained for the 12 conditions in which standard error termination rule is used (see Table 1) shows that under all conditions, the correlation for θ_g was high, while the correlations for θ_2 and θ_3 were medium-low, with all being significant. While error statistics were relatively low for the general trait, they were high for specific factors. In instances where the D-rule method was used, the correlation obtained for specific factors was higher than with other methods, while the error estimations were lower. When weighting was used in item selection methods, the weighting improved the estimations obtained for θ_g as expected, while causing a drop in the values obtained for specific factors. A significant reduction in the average number of items administered (k) was observed, especially when weighting was used in the W-rule method. When the ability estimation method is being accounted for, the average number of items administered is much lower in instances using MAP compared to those using EAP. Therefore, it may be stated that MAP generally shows higher performance than EAP. The high correlations and the low standard error rates obtained for the general trait may be explained as part of the nature of bifactor structure. This is supported by the fact that one of the fundamental characteristics of the bifactor model is its explanatory power for a large portion of the variance in the variable through the general trait, while a small portion is explained by the specific factors (Reise, 2012). Therefore, it may be stated that estimations obtained for the general trait are expected to be more in line with the estimations obtained from PPTs rather than specific factors.

Regarding the faultlessness of the estimations obtained for the general trait within the framework of the standard error termination rule, all item selection methods portrayed similar performance, and weighting methods reduced the number of items administered as expected. Additionally, all other item selection methods had lower performance on specific factors compared to the D-rule method. As such, it was concluded that for the standard error termination rule, MAP ability estimation and the D-rule item selection method was the condition with the highest performance.

Following the determination of the best condition among the 12 using the standard error termination rule, the conditions based on θ convergence ($\Delta\theta < .05$) termination rule were evaluated. The results (see Table 2) obtained with MAP were found to be better than all of the item selection methods obtained with EAP. While the correlation and error statistics obtained for the general trait were similar for all the item selection methods, D-rule was found to provide the best results for specific factors once again. Regarding number of items administered, D-rule resulted in the highest values while the lowest were obtained when weighting was applied for the general trait. Despite the fact that the number of items administered to is relatively higher with the D-rule method, it portrays similar performance with other methods regarding the general trait and much better performance regarding specific factors. This led to the conclusion that the D-rule item selection method was optimal for conditions in which the θ convergence termination rule is used.

Lastly, the values obtained for the 12 conditions within the scope of the fixed number of items termination rule (k = 20) were reported (see Table 3). As with the other 24 conditions, the results show similar levels with all item selection methods of the estimated correlation and error values for the general trait in the 12 conditions where a fixed number of items termination rule is applied. The results obtained for specific factors also had high performance when the D-rule item selection method was used. The performance it provides regarding the general trait is at a similar level to other item selections and higher than them on specific factors. This resulted in the determination that use of the D-rule item selection method in conditions with a fixed number of items termination rule was more suitable, and that the results obtained with MAP were slightly better than those of EAP, concluding that this method is preferable for ability estimation.

Table 2. *Correlation, bias, RMSD, standard error values and avarage number of items administered for conditions using standard error termination rule*

Termination Rule	Ability Estimation Method	Item Selection Method	r			bias			RMSD			SE			
			$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	k
Standard Error (SE < .4)	EAP	D-rule	.91	.63	.66	-.0441	.0472	.0454	.38	.55	.54	.37	.69	.66	16.1
		KL	.92	.41	.58	-.0029	-.0037	-.0288	.35	.61	.55	.38	.87	.77	17.7
		W-rule	.93	.41	.58	-.0013	-.0030	.0293	.35	.61	.55	.37	.87	.78	17.8
		W-rule (.8,.1,.1)	.93	.33	.47	-.0074	.0115	.0200	.34	.62	.59	.35	.91	.83	12.9
		T-rule	.93	.44	.57	-.0273	.0249	.0476	.35	.60	.58	.35	.86	.71	13.0
		T-rule (.8,.1,.1)	.93	.35	.51	-.0081	.0187	.0328	.35	.62	.60	.34	.90	.78	12.7
	MAP	D-rule	.90	.60	.62	-.0326	.0370	.0721	.39	.56	.54	.39	.71	.65	13.4
		KL	.92	.41	.56	.0012	-.0007	.0019	.36	.61	.55	.39	.87	.78	15.2
		W-rule	.92	.40	.57	.0026	.0018	.0027	.36	.61	.55	.39	.87	.78	15.2
		W-rule (.8,.1,.1)	.92	.30	.46	.0079	.0064	.0329	.35	.62	.58	.36	.92	.84	11.4
		T-rule	.92	.42	.55	-.0195	.0084	.0806	.35	.60	.57	.36	.86	.69	11.5
		T-rule (.8,.1,.1)	.92	.33	.49	.0037	.0050	.0482	.35	.62	.58	.35	.90	.77	11.33

Table 3. *Correlation, bias, RMSD, standard error values and avarage number of items administered for conditions using ϑ convergence termination rule*

Termination Rule	Ability Estimation Method	Item Selection Method	r			bias			RMSD			SE			
			$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	k
θ Convergence (Δθ < .05)	EAP	D-rule	.94	.71	.71	-.0301	-.0678	.0515	.32	.53	.50	.30	.63	.62	24.1
		KL	.92	.42	.59	.0073	.0006	-.0289	.35	.60	.54	.37	.87	.77	18.1
		W-rule	.92	.41	.59	.0113	.0043	-.0282	.35	.61	.54	.37	.88	.77	18.0
		W-rule (.8,.1,.1)	.94	.39	.53	-.0031	.0173	.0152	.32	.62	.57	.31	.88	.79	16.1
		T-rule	.94	.50	.62	-.0164	.0230	.0473	.32	.59	.55	.30	.80	.68	17.1
		T-rule (.8,.1,.1)	.94	.41	.55	-.0095	.0216	.0337	.32	.62	.58	.30	.86	.75	16.3
	MAP	D-rule	.94	.70	.70	-.0069	.0639	.0768	.32	.52	.49	.33	.63	.61	21.4
		KL	.92	.41	.59	.0208	.0013	-.0072	.36	.60	.54	.38	.87	.76	17.2
		W-rule	.92	.40	.59	.0218	.0037	-.0032	.36	.61	.54	.38	.87	.76	17.1
		W-rule (.8,.1,.1)	.94	.36	.53	.0108	.0078	.0320	.32	.62	.56	.32	.88	.77	15.1
		T-rule	.94	.50	.63	.0030	.0086	.0684	.32	.59	.53	.31	.79	.65	16.6
		T-rule (.8,.1,.1)	.94	.42	.53	.0088	.0055	.0570	.32	.61	.57	.32	.85	.74	15.0

Table 4. *Correlation, bias, RMSD and standard error values for conditions using fixed number of items termination rule*

Termination Rule	Ability Estimation Method	Item Selection Method	r			bias			RMSD			SE		
			$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$
Fixed Number of Items (k = 20)	EAP	D-rule	.93	.69	.69	-.0315	.0637	.0481	.33	.53	.51	.34	.65	.63
		KL	.93	.42	.60	.0085	-.0023	-.0344	.33	.61	.54	.37	.87	.76
		W-rule	.93	.42	.61	.0108	-.0012	-.0336	.33	.61	.53	.35	.87	76
		W-rule (.8,.1,.1)	94	.42	.60	-.0093	.0183	.0115	.30	.62	.55	.29	.85	.72
		T-rule	.95	.53	.66	-.0154	.0277	.0346	.30	.58	.53	.29	.78	.66
		T-rule (.8,.1,.1)	.95	.50	.58	-.0097	.0231	.0266	.30	.59	.57	.28	.81	.72
	MAP	D-rule	.93	.69	.69	-.0093	.0607	.0751	.33	.52	.50	.33	.64	.61
		KL	.93	.43	.61	.0170	-.0054	-.0086	.33	.60	.53	.35	.86	.75
		W-rule	.93	.42	.61	.0184	.0009	-.0061	.33	.61	.53	.35	.86	.75
		W-rule (.8,.1,.1)	.95	.42	.61	.0078	.0071	.0255	.29	.62	.53	.30	.84	.70
		T-rule	.95	.53	.66	-.0027	.0144	.0601	.29	.58	.52	.29	.76	.64
		T-rule (.8,.1,.1)	.95	.50	.58	.0037	.0124	.0495	.29	.59	.55	.29	.80	.80

To determine the most suitable condition for use in the real-time MCAT as a result of the simulations, a final evaluation was conducted for the three conditions with the best results for all termination rules. The error rates and correlation statistics of these conditions are provided in unison (see Table 4), and the lavaan (Sarkar, 2016) package in R was used to graph each one individually (see Figure 4), with the best condition for the real-time MCAT application being decided as a result of these values and graphs.

For three different termination rules, the best results were obtained using D-rule item selection and MAP ability estimation methods. Of these three conditions, the one with the highest measurement accuracy for the real-time application was determined by studying the graphs obtained for the general trait.

Table 5. *Statistics of the best conditions for each termination rule*

Termination Rule	Ability Est. Method	Item Selection Method	r			bias			RMSD			SE			k
			$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	
Standard Error	MAP	D-rule	.90	.60	.62	-.033	.037	.072	.39	.56	.54	.39	.71	.65	13.4
θ Convergence	MAP	D-rule	.94	.70	.70	-.007	.064	.077	.32	.52	.49	.33	.63	.61	21.4
Fixed Number	MAP	D-rule	.93	.69	.69	-.009	.061	.075	.33	.52	.50	.33	.64	.61	20

Firstly, the standard error - θ_g graph for the three conditions was obtained for the general trait. In this case, as termination is based on .4 standard error, despite only the maximum (60) number of items administered, the estimations that don't fall below this standard error value are still above .4. It is notable that these high standard error values are observed with individuals with high θ levels.

The second and third graphs were obtained for θ convergence and for fixed number of items termination rules, and these graphs appear similar to 3 each other. Both graphs have a very small range for standard error values towards the center of the ability scale. However, as the estimated θ value of examinees increases, the standard error value increases and the values obtained go as high as .6. This situation may stem from the fact that the medium level ability estimations (θ = 0) of the item pool provide more information, while anything beyond θ = 1 provides less information. It is also notable that the standard error value obtained with θ convergence disperses over a wider range compared to that obtained with fixed number convergence.

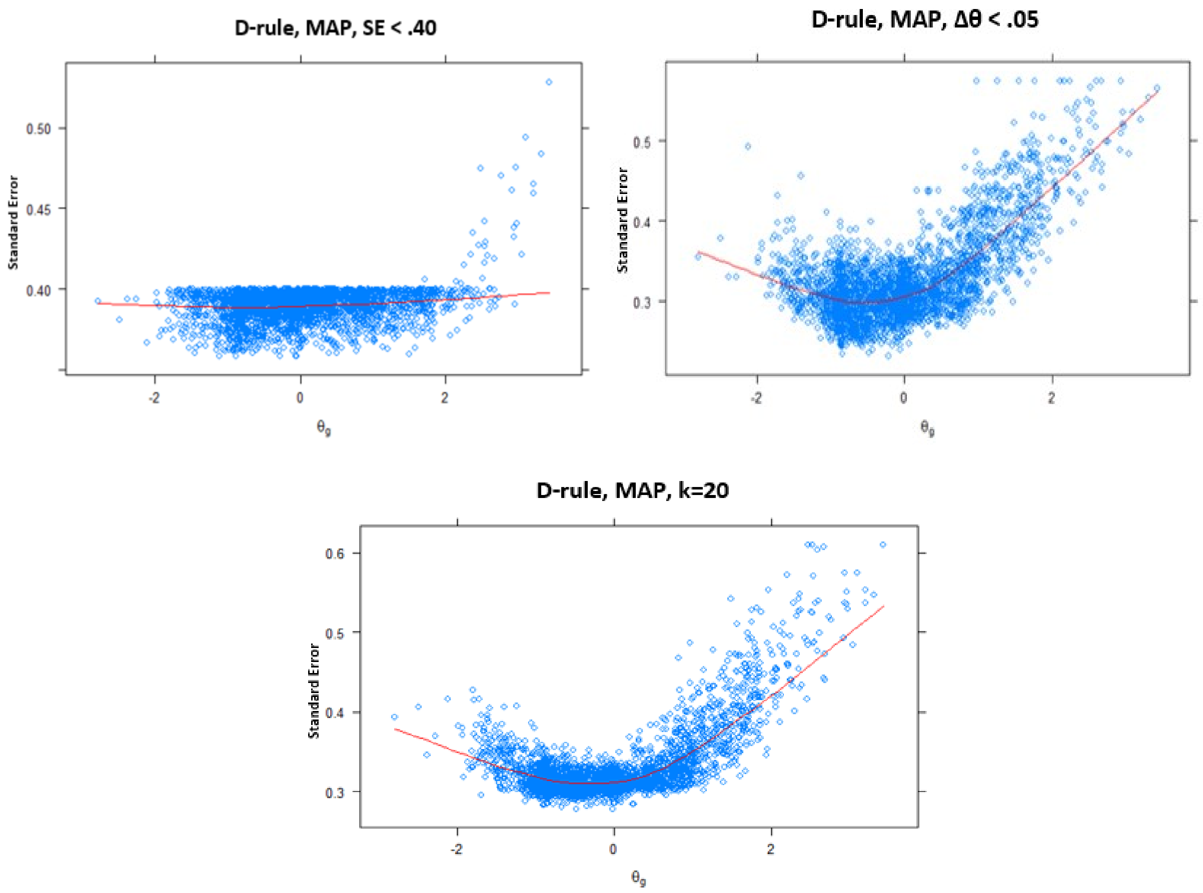


Figure 4. *Standard error – θ_g graphs*

When the number of items administered – θ_g graphs obtained for variable length applications are studied (see Figure 5), in cases where the standard error termination rule is used; the average number of items administered is near 10 throughout a large portion of the θ scale, and this value increases as θ approaches 2. It is observed that individuals with high ability levels reached the maximum number of items to be administered, in addition to the termination rule. In the condition where the θ convergence termination rule is used, it is notable that the average number of items administered over the whole ability scale has a high and wide range.

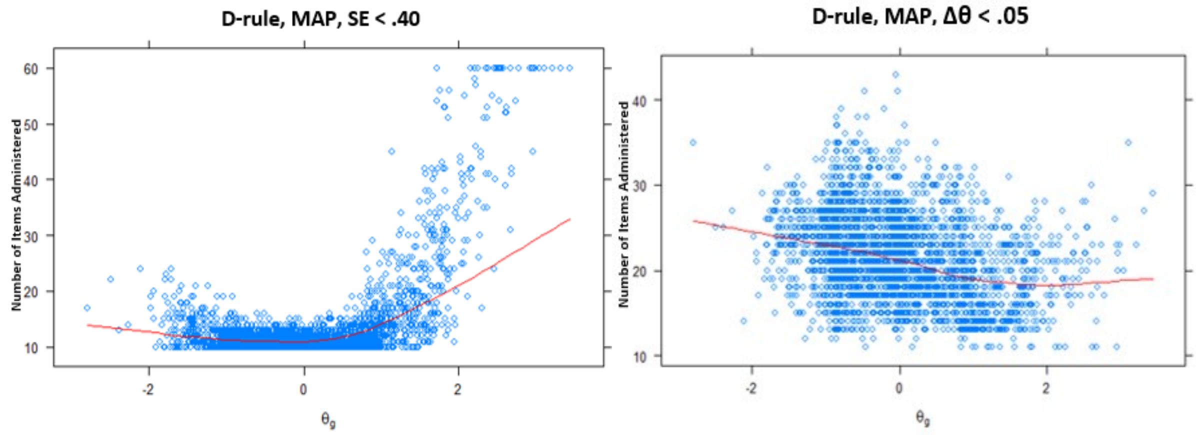


Figure 5. *Number of items administered - θ_g graphs*

In conditions where standard error termination rule in the hybrid simulation are used and the frequency values of the item numbers responded are studied (see Figure 6), it is notable that approximately 30% of the 3057 participants responded to 10 items, the minimum determined to terminate the test. Additionally, based on this graph, it may be stated that approximately 85% of the individuals responded to 10-15 items.

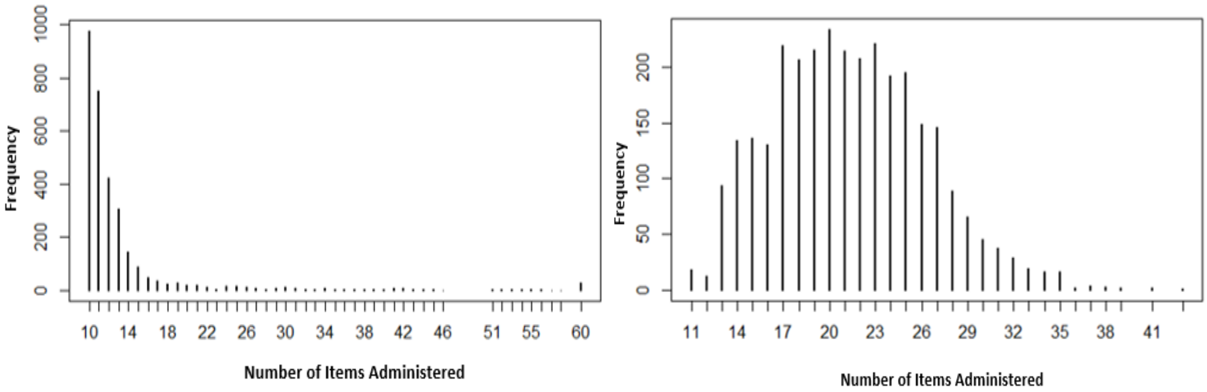


Figure 6. *Frequency of number of items administered for standard error & θ convergence termination Rules*

In a large number of participants, the number of items administered in the condition using the θ convergence termination rule varied between 17-25. The difference from the condition using the standard error termination rule is that the range of the number of items answered in this condition is narrower.

Based on these graphs, it may be stated that the use of the standard error termination rule in the real-time application is more efficient than other methods regarding number of items administered. After a comparison of the graphs and statistics obtained for the best conditions, as a result of the simulations for each termination rule; MAP was selected as the ability estimation method, D-rule was selected as the item selection method, and standard error as a termination rule (.4) was selected as the most appropriate components of the algorithm for the real-time MCAT application.

3.2. Real-Time MCAT Application Results

Based on the results obtained for 36 different conditions regarding the hybrid simulation, D-rule was chosen as the item selection method, MAP as the ability estimation method, and a .40 value cutoff in the standard error for the general trait as a termination rule was decided on during the real-time MCAT application. In addition, a minimum of 10 items to be administered to each participant for the test termination, and test termination after 60 items in instances where standard error remained above .40 criteria were applied. Based on this algorithm, the real-time MCAT application was conducted using the mirtCAT (Chalmers, 2016) package and the shiny (Chang, 2019) GUI package for R on 99 students in the final semester of the preparatory class.

Studying the frequencies of the number of items administered (see Table 6) shows that 74 participants answered 10-12 items. 12 participants answered 13 items, while the number of participants who responded to 14 items was 4, and 15 items was 5. Only 4 participants answered more than 15 items.

The results obtained show that the average number of items administered to the 99 students participating in the real-time application is 12.3. This value is close to but slightly lower than the average number of items of 13.4 obtained during the simulation application using the same condition (D-rule, MAP, SE<.4) as the real-time application. The number of items administered varies between 10 and 45. Regarding the grammar and vocabulary skills measured by the real-time application, the number of items examinees answered in the PPT is 50. This led to the conclusion that in the real-time MCAT application, examinees are administered an average of 74.4% fewer items than the PPT.

Table 6. *Distribution of number of items administered during the real-time application*

Number of Items Administered	Frequency	%
10	25	25.3
11	24	24.2
12	25	25.3
13	12	12.1
14	4	4.0
15	5	5.1
>15	4	4.0
Total	99	100.0

Following the real-time MCAT application, it was observed that 60 of the items in the 200 present in the item pool were used, while 140 were not present in any of the applications. In other words, in the real-time MCAT application conducted with 99 individuals, 30% of the item pool was used. Of the 60 items used, it is notable that 37 of them have used numbers under 5.

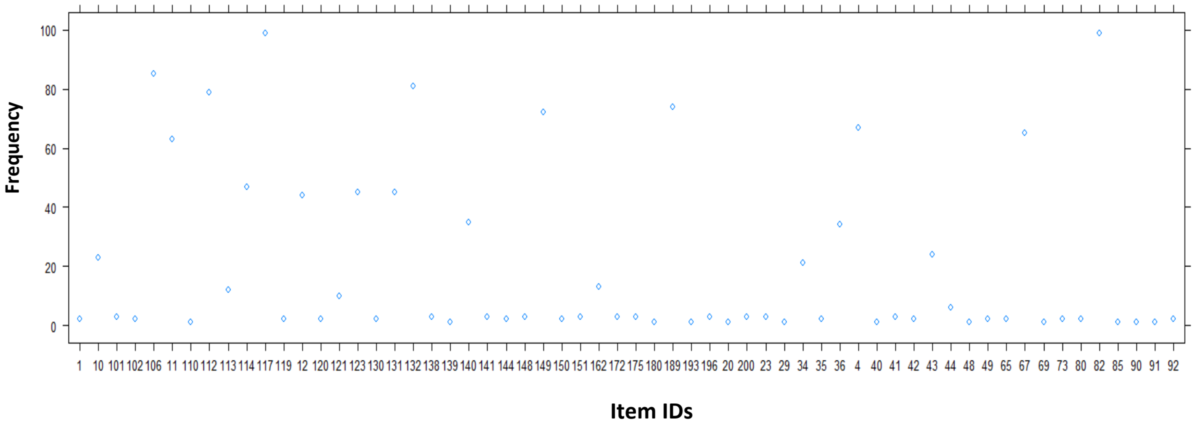


Figure 7. *Item use from the item pool*

When the use frequencies of the 60 items from the item pool used at least once in the real-time application are studied (see Figure 7), it was found that of these items, item number 117 was used at the beginning of every application, and the 82nd item was present in all the applications. Other than these two items, 8 items were administered at least in 60 applications. The number of items administered 20 or more times was 19.

The real-time MCAT application was conducted on 99 students studying at an English preparatory class at a university in Turkey. Of these students, 32 entered their proficiency examination one month before the application was conducted. Through this opportunity, the correlation between the real-time MCAT application and their PPT results (proficiency examination) were calculated. These calculations resulted in a .77 correlation between the general trait estimations resulting from the real-time MCAT application and their total score obtained from the PPT.

4. DISCUSSION and CONCLUSION

Within the scope of this study, grammar and vocabulary data of English preparatory class students were gathered from their proficiency examinations required to attend undergraduate courses, and an MCAT measuring the general trait and their grammar and vocabulary was developed. To this end, an item pool consisting of four separate groups with 50 items each was established. This was followed by a hybrid simulation application to determine the algorithm to be used in the real-time MCAT application. Following this simulation, the ability estimation methods (EAP and MAP), item selection methods (D-rule, KL, W-rule, T-rule, weighted W-rule and weighted T-rule), and the termination rules (standard error, θ convergence and fixed number of items) were used to create 36 different conditions. For each dimension in these conditions, the correlation between the real and estimated θ values, bias, RMSD and standard error values were obtained. Due to the fact that in addition to correlation values and error statistics, the average number of items administered is also an important indicator of measurement accuracy in CAT applications, the number of items administered in the conditions using termination rules based on variable test length were also reported. Following the determination of the most appropriate MCAT algorithm based on the simulations for the real-time application, this algorithm was used to conduct the real-time MCAT application. The correlation between the PPT scores and MCAT real time application results of the 32 examinees was also calculated. Additionally, the item use frequencies of items in the pool and number of items administered to each 99 examinee participating in the application were reported. In this section, the results are presented under separate headings for the hybrid simulation and the real-time MCAT application.

4.1. Interpretation of Findings Obtained from the Hybrid Simulation

The simulation results indicate that for the three termination rules used within the scope of the study, the most appropriate conditions the ones where D-rule item selection and MAP ability estimation methods used. While the D-rule item selection method provided similar performance with other methods regarding general ability estimation, it provided much better values than other methods for specific factors. These findings are similar to the study of Seo and Weiss (2015), who suggested the use of D-rule item selection and MAP ability estimation methods in situations where estimations for specific factors are important.

Within the scope of this study, the correlations between the real and estimated ability parameters for the general ability were quite high under all conditions, while the correlations for specific factors were lower. The first reason for this may be the nature of the bifactor model. This is due to the known given that for a multidimensional model to fit with a bifactor structure, the structure must not only estimate a general ability but the factor loadings for the general ability must be higher than group factors (Reise, Morizot & Hays, 2007). It may be stated that this situation causes a reduction in given information for specific factors as a structure adapts to the bifactor model. In addition, Seo (2011) found that similar to this study, the correlation values obtained for the general ability are higher than those obtained for specific factors.

While the correlation values obtained for the general ability were high, another reason the values for group factors being low is the number of items administered. Weiss and Gibbons (2007) indicate that to increase the efficiency of bifactor MCATs, between 20 and 50 items must be used for each specific factor. Other related studies in the literature on bifactor MCAT applications such as Seo (2011) and Seo and Weiss (2015) also used 20 items for each group factor. In Sunderland et al.'s (2019) study, which aimed to estimate internalizing through a bifactor MCAT application, it was found that a 133 item PPT scale was completed in an average of 44 items. Nieto, Abad and Olea (2018) developed an MCAT application based on a bifactor model of the big five scale, and concluded that a result was obtained for each dimension through 12 items on average. Within the scope of this study, 10 items were used for each factor in the fixed number of items termination rule condition, while the number of items answered fell as low as 5 for each specific factor for a large portion of the other conditions. This may be the cause of the low correlations obtained for specific factors and the high error statistics. Within the scope of this study, the researcher aimed to develop a real-time application for a 50-item PPT application. The number of items administered in CAT studies is directly related to measurement accuracy. Additionally, considering the real-time application of this study aims to obtain an overall score estimation without disregarding multidimensionality, it was predicted that determining the minimum number of items to be answered for each dimension as 20 would reduce the efficiency of the real-time application.

When the item selection methods with weighting were studied, within termination rules based on variable test length, use of W-Rule item selection methods with weighting results in a rise in the correlations for general ability and a significant reduction in error statistics. The number of items answered with weighting was reduced by 20-25% on average. Additionally, the improvement in the performance of T-rule with weighting was higher than with W-rule.

In applications using the standard error termination rule, especially with ability levels where the item pool information level is low, the estimated standard error levels were observed to be high. As such, in applications where the item pool is not large enough, it may be stated that the use of a standard error termination rule is more appropriate.

4.2. Interpretation of Real-Time MCAT Application Findings

Following the real-time MCAT application, 30% of the 200 item pool was used. Based on these values, it may be stated that the use rate of the pool is low. However, studies indicate that in 50% of CAT applications, only 14% of the item pool is used (Wainer, 2000). Considering the item pool information level is high for a mid-low θ level, middle or low ability levels of the examinees may be causing the use of only a small portion of the pool. In addition, as the students who participated in the real-time MCAT application are from the same course level and therefore at similar ability levels regarding the test, the high use rate of certain items is to be expected. In such instances, the use of very easy and very difficult items are expected to be low (Wei & Lin, 2015). Additionally, the average number of items administered being low at 12.3 and the lack of an item exposure control method within the scope of the study may also be causes behind the low use rates of the item pool. The generally similar ability levels and the lack of an item exposure rate control mechanism leads to the conclusion that the limited number of items administered are frequently the same items.

Of the 60 items used in real-time MCAT application, only 23 were included in 5 or more of the 99 applications. This situation is similar to Veldkamp and van der Linden’s (2002) MCAT study in which over 80% of the tests only used 20% of the item pool. As stated earlier, as the average number of the items administered is low and 75% of the examinees responding to fewer than the average number of items administered may have been effective in this situation emerging.

The findings show that the real-time MCAT application lasts approximately 9 minutes. When considering each question is allocated one minute in a PPT, it may be stated that the MCAT application takes 80% less time than PPT. This is considered to be important regarding the effectiveness of CAT applications.

4.3. Recommendations for Future Research

Within the scope of this study, while conducting a general ability estimation based on a common source of variance for all items, a bifactor model that takes into account multidimensionality was used. It may be stated that bifactor models are an alternative to high-order/hierarchical models (Seo & Weiss, 2015). The only MCAT study in the literature using high-order IRT models was conducted by Huang, Chen and Wang (2012). It is believed that a comparison between the findings of this current study and an MCAT study using high order IRT models would contribute significantly to the literature.

Despite the increase in the number of studies in the recent years on bifactor MCAT, the literature in this field is still limited. Some of these studies were conducted beyond the scope of the purpose of this study (Zheng et al., 2013; Gibbons et al., 2016). It is also noted that all of the applications conducted regarding real-time applications in bifactor MCAT studies aim to study affective characteristics. It may therefore be stated that a need has arisen for bifactor MCAT applications with different situations in which the aim is to produce an overall score for a multidimensional cognitive ability – as with this study.

The item pool used within the scope of this study is limited. Future research may use item exposure control methods in real-time applications based on larger item pools, and these methods may allow a comparison between the performance of item selection and ability estimation methods.

Within this study, only items with independent response status were used. However, the measurement of language skills also requires applications based on the response to more than one item based on a text, image, etc. As such, it is documented that testlet-based IRT models may be used (e.g. Frey, Seitz & Brandt, 2016). Additionally, the bifactor model used within the scope of this study may be used for testlet-based tests (see DeMars, 2006). Thus, it is believed that an MCAT application based on the bifactor model for tests of skills beyond the scope of this study such as reading and listening, which are some of the fundamental dimensions of language skills, would contribute to the research.

Acknowledgements

This study presents partial findings of the doctoral dissertation entitiled *Examining the Results of Multidimensional Computerized Adaptive Testing Applications in Real and Generated Data Sets* (Şahin, 2017).

Declaration of Conflicting Interests and Ethics

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

ORCID

Murat Doğan ŞAHİN https://orcid.org/0000-0002-2174-8443
Selahattin GELBAL https://orcid.org/0000-0001-5181-7262

5. REFERENCES

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

Bulut, O & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian Journal of Educational Research, 49*, 61-80.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algortihm. *Pschometrika, 46*(4), 443-459.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*(5), 1-38.

Chang, W. (2019). Shiny: Web application framework for R. Version 1.3.2

Choi, S. W., Grady, M. W., & Dodd, B. G. (2010). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement, 71*, 37-53.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement. 43*(2), 145–168.

Eggen, T. (2007). *Choices in CAT models in the context of educational testing*. Paper presented at the CAT Models and Monitoring Paper Session, June 7, 2007 (Retrieved November 11, 2016, from http://publicdocs.iacat.org/cat2010/cat07eggen.pdf).

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

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-374.

Ferrando, P. & Chico, E. (2007). The external validity of scores based on the twoparameter logistic model: Some comparisons between IRT and CTT. *Psicológica, 28*, 237-257.

Frey, A. & Nicki-Nils, S. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation, 35*. 89-94

Frey A, Seitz N-N and Brandt S (2016) Testlet-Based Multidimensional Adaptive Testing. *Front. Psychol., 7*, 1758.

Gelbal, S. (1994). *P madde güçlük indeksi ile Rasch modelinin b parametresi ve bunlara dayalı yetenek ölçüleri üzerine bir karşılaştırma*. Unpublished doctoral dissertation. Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology, 12*, 83-104.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*(4), 49-58.

Gustafsson, J., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28*, 407-434.

Hambleton, R. K. & Swaminathan, H. (1985*). Item response theory: Principles and applications.* Boston, MA: Kluwer Academic Publishers.

Huang, H., Chen, P & Wang, W. (2012). Computerized adaptive testing using a class of high-order item response theory. *Applied Psychological Measurement, 36*(8), 689-706.

Huebner, A. R., Wang, C., Quinlan, K. & Seuber, L. (2016). Item exposure control for multidimensional computer adaptive testing under maximum likelihood and expected a posteriori estimation. *Behav. Res., 48*, 1443-1453

Jabrayilov, R., Emons, W. H. M. & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement, 40*(8) 559-572.

Kalender, I., & Berberoglu, G. (2017). Can computerized adaptive testing work in students' admission to higher education programs in Turkey? *Educational Sciences: Theory & Practice, 17*, 573-596.

Kelecioğlu, H. (2001). Örtük özellikler teorisindeki b ve a parametreleri ile klasik test teorisindeki p ve r istatistikleri arasındaki ilişki, *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 20*, 104-110.

Lawson, S. (1991). *One parameter latent trait measurement: Do the results justify the effort?* In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments. Greenwich, CT: JAI.

Li, Y. H., & Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement, 29*, 3-25.

Lin, C. & Chang, H. (2018). Item Selection Criteria with Practical Constraints in Cognitive Diagnostic Computerized Adaptive Testing. *Educational and Psychological Measurement, 79*(2), 335-357.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*(4), 389-404.

Ndalichako, J. L., & Rogers, W. T. (1997). Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. *Educational and Psychological Measurement, 57*, 580-589.

Nieto, M. D., Abad, F. J., & Olea, J. (2018). Assessing the Big Five with bifactor computerized adaptive testing. *Psychological Assessment, 30*(12), 1678-1690.

Nydick, S. & Weiss, D. J. (2009). *A hybrid simulation procedure for developments of CATs*. Paper presented at the Item Pool Development Paper session at the 2009 GMAC Conference on Computerized Adaptive Testing.

Progar, S. & Sočan, G. (2008). An empirical comparison of Item Response Theory and Classical Test Theory. *Horizons of Psychology, 17*(3), 5-24.

Reckase, M., D. (2009). *Multidimensional item response theory: Statistics for social and behavioral sciences*. New York, NY: Springer.

Reise, S. P. (2012). The rediscovery of bifactor measurement models, *Multivariate Behavioral Research, 47*(5), 667-696.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19-31.

Sarkar, D. (2016). *Lattice: Multivariate Data Visualization with R*. Springer.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331-354.

Segall, D. O. (2005). *Computerized adaptive testing.* In K. Kempf-Leonard (Ed.), Encyclopedia of Social Measurement. New York: Academic Press.

Seo, D. G. (2011). *Application of the bifactor model to computerized adaptive testing*. Unpublished Doctoral Disertation. University of Minnesota.

Seo, D. G. & Weiss, D. J. (2015). Best Design for Multidimensional Adaptive Testing with the Bifactor Model. *Educational and Psychological Measurement, 75*(6), 954-978.

Su, Y. (2016). A comparison of constrained item selection methods in multidimensional computerized adaptive testing. *Applied Psychological Measurement, 40*(5) 346-360.

Sunderland, M., Batterham, P. Carragher, N., Calear, A. & Slade, T. (2019). Developing and Validating a Computerized Adaptive Test to Measure Broad and Specific Factors of Internalizing in Community Sample. *Assessment, 26*(6) 1030-1045.

Şahin, M. D. (2017). *Examining the Results of Multidimensional Computerized Adaptive Testing Applications in Real and Generated Data Sets* [Unpublished doctoral dissertation]. Hacettepe University, Graduate School of Educational Sciences, Ankara.

Thompson, N. A. & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*(1), 1-9.

van der Linden, W. J. (2016). *Handbook of Item Resonse Theory*. Boca Raton: CRC Press.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*, 575-588.

Wang, C., Chang, H. & Boughton, K. A. (2012). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement, 37*(2), 99-122.

Wainer, H. W., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L. & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbaum.

Wainer, H. (2000). Rescuing computerized testing by breaking Zipf's law. *Journal of Educational and Behavioral Statistics, 25,* 203-224.

Ware J. E., Kosinski, M., Bjorner, J. B., Bayliss, M. S., Batenhorst, A., Dahlo, C. G. H., Tepper, S. & Dowson, A. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research, 12*, 935-952.

Wei, H., & Lin, J. (2015). Using out-of-level items in computerized adaptive testing. *International Journal of Testing, 15*(1), 50-70.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*, 774-789.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2), 70-84.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 1-27.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measuremen, 21*(4), 361-375.

Weiss, D. J., & Gibbons, R. D. (2007). *Computerized adaptive testing with the bifactor model*. Paper presented at the New CAT Models session at the 2007 GMAC Conference on Computerized Adaptive Testing. Retrieved-October 12, 2016, from http://publicdocs.iacat.org/cat2010/cat07weiss&gibbons.pdf

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and Applications. *Psychometrika, 77*, 495-523.

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement, 37*, 3-23.

Yao, L. (2014). Multidimensional CAT item selection methods for domain scores and composite scores with item exposure control and content constraints. *Journal of Educational Measurement, 51*, 18-38.

Yao, L., Pommerich, M & Segall, D. O. (2014). Using multidimensional CAT to administer a short, yet price, screening test. *Applied Psychological Measurement, 38*(8) 614-631.

Zheng, Y., Chang, C. H., & Chang, H. H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 22*, 491-499.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 343–360

https://doi.org/10.21449/ijate.679876

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

Reliability and Validity of TPACK Instruments in EFL

Abdullah Arslan[1,*]

[1]Shanghai International Studies University, Shanghai, China

ARTICLE HISTORY

Received: Jan. 25, 2020
Revised: July 3, 2020
Accepted: July 29, 2020

KEYWORDS

Reliability,
Validity,
TPACK,
English teachers

Abstract: In this study, issues of validity and reliability of a wide range of instruments used to measure technological pedagogical content knowledge level of English teachers were discussed. To this end, the search in the databases of ERIC, ScienceDirect, Scopus, EBSCOhost, and Web of Science was conducted. As a result of applying a set of criteria to publications retrieved from the databases, 60 studies (including 40 articles, 14 dissertations and 6 conference papers) were found suitable for analysis in the current study. A two-level analysis was conducted. First one was study-level analysis focusing on general characteristics of each study and the second one was an instrument-level analysis that focuses on target audience and research instruments. As a consequence of the analysis at the study-level, 128 instruments were classified into five types of instruments including, open-ended questionnaire, observation, performance assessment, interview, and self-report instruments. At the instrument-level analysis, issues of validity and reliability of those instruments and target audience were investigated. The findings revealed that 60% of the reviewed studies did not provide any index of reliability, and similarly over 80% of the studies presented no evidence of validity.

1. INTRODUCTION

The advances of innovative technology have paved the way for the emergence of the concept of educational technology. Educational technology is fundamentally composed of some components that are constantly interrelated to each other. The design and development of educational content is closely related to its employment and management. Besides, one of the indispensable and crucial components of this process is the assessment of educational content in terms of both students' learning efficiency and effectiveness of materials (Luppicini, 2005). Educational technology has been the focus of different stakeholders' attention in education since there have been great efforts of nations to pursue the integration of technology with education approaches (Chai et al., 2013). Despite the technological developments, there are some concerns among some scholars whether teachers could use technological tools as they are meant to instead of merely supporting traditionally oriented teaching (Agyei &Voogt, 2012; Shin et al., 2009; Sessoms, 2008). At this point, in order to assess how teachers are able to integrate their knowledge of content, pedagogy and technology, technological pedagogical content knowledge (TPACK, hereafter) steps in (Koehler & Mishra, 2009; Koh et al., 2013; Schmidt et al., 2009).

TPACK contains three mutually interconnected knowledge domains. These domains are called

CONTACT: Abdullah ARSLAN ✉ abdullahaaarslan@gmail.com Shanghai International Studies University, Shanghai, China

ISSN-e: 2148-7456 /© IJATE 2020

as content knowledge (CK), pedagogical knowledge (PK), and technological knowledge (TK) where teachers are supposed to integrate their content, pedagogy, and technology knowledge into their teaching process to accomplish efficient and effective learning process on students' parts (Drajati et al., 2018). Technology Knowledge (TK) is simply defined as the knowledge of operating computer software and hardware and employment of a range of software like presentation slides, spreadsheet program, word processors, and some tools for communication. Moreover, teachers are expected to have the ability to run above-mentioned tools and technologies, and use them effectively in the process of teaching (Chai et al., 2010; Mishra & Koehler, 2006, 2008). Content Knowledge (CK) refers to knowledge of teachers' subject area where they are supposed to have a good command of expressing and explaining fundamental facts of the content knowledge, concepts, theories, and protocols. Furthermore, they are expected to have the ability to connect ideas with each other by evaluating knowledge of the content (Chai et al., 2010; Mishra & Koehler, 2006, 2008). As for Pedagogical Knowledge (PK), it refers to the strategies, methods, or tactics teachers employ in teaching process where they should be responsible for planning, implementing, managing, and evaluating educational activities of students with an effort to specify and assess how students acquire skills and construct their knowledge through cognitive and social constructivism approaches in classroom environment (Mishra & Koehler, 2006, 2008).

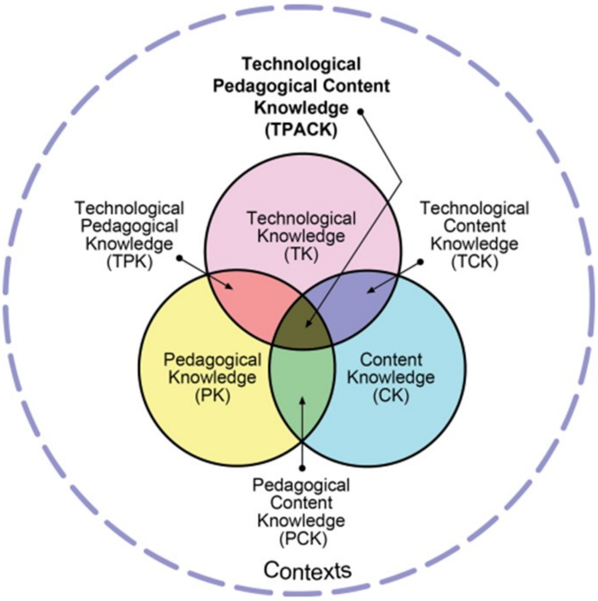

Technological Pedagogical Content Knowledge (TPACK)
Technological Pedagogical Knowledge (TPK)
Technological Knowledge (TK)
Technological Content Knowledge (TCK)
Pedagogical Knowledge (PK)
Content Knowledge (CK)
Pedagogical Content Knowledge (PCK)
Contexts

Figure 1. *Technological pedagogical content knowledge framework* (source: Koehler & Mishra, 2008)

Furthermore, there are four other domains emerged from the intersection of aforementioned three knowledge domains (Figure 1). These domains are named as Technological Pedagogical Knowledge (TPK), Pedagogical Content Knowledge (PCK), Technological Content Knowledge (TCK), and TPACK. The first coalescence is comprised of technological pedagogical knowledge (TPK) that involves the bond between technologies and pedagogical practices. The second one is pedagogical content knowledge (PCK), which is directly related to pedagogical practices and learning objectives. The third one is technological content knowledge (TCK) that expresses the relation between technologies and learning objectives. Being composed of the intersection of the aforementioned coalescence that displays a very complicated relation between the areas of knowledge, TPACK is generally defined as a dynamic conceptual framework teachers may use to design and deliver course content by employing technology to facilitate and enhance student learning process (Graham, 2011; Niess, 2011). It is also regarded as an instrument that assesses and reflects teachers' skills to combine pedagogy, content and technology flexibly with their act of teaching (Harris et al., 2010; Schmidt et al.; 2009; Mishra & Koehler, 2006). In the field of education, TPACK has been the focal centre of researchers' interests. To illustrate, some researchers use TPACK as a self-assessment or self-reporting instrument to measure teachers' efficacy (Jen et al., 2016; Koh & Divaharan, 2013; Mouza et al., 2014; Schmidt et al., 2009; Tschannen-Moran & Hoy 2001). In addition, a body of research has made an attempt to both investigate artefacts designed by teachers (Harris et al., 2010; Koh et al., 2013) and explore teachers' performances through TPACK-based educational technology courses and activities (Graham et al., 2012; Jang & Tsai, 2012; Kafyulilo et al., 2015; Kramarski&Michalsky, 2010; Tokmak et al., 2013). In some studies, quite a few instruments are designed for the measurement of TPACK in specific areas such as science teachers (Canbazoglu-Bilici et al., 2013), geography teachers (Su et al., 2017), mathematics teachers (Bowers & Stephens, 2011), and language teachers (Baser et al., 2016; Chai et al., 2013).

In the field of EFL, the literature reveals that researchers are generally inclined to employ TPACK as a self-reporting instrument to assess perceptions, self-efficacy, competency, and skills of teachers. For example, in order to evaluate the effectiveness of intervention on TPACK in a qualitative study, Koçoğlu (2009) investigates how pre-service English teachers improve technology integration into their teaching practice. The study concludes that pre-service English teachers acquire high TPACK skills. In the same way, Kurt et al. (2014) examine Turkish pre-service English teachers' TPACK development in a 12-week intervention based on Learning Technology by Design approach (Mishra & Koehler, 2006) through the survey of Pre-service Teachers' Knowledge of Teaching and Technology (Schmidt et al., 2009). The results of the study report that there is a statistically significant increase in participants' TK, TCK, TPK and TPACK scores. In a mixed-method design Ersanlı (2016) questions the effectiveness of five-week training of pre-service English teachers. In the study, data are collected through TPACK Competency Survey (Archambault & Crippen, 2009) and journal entries of the participants. The results reveal that there is a statistically significant improvement in participants' TPACK scores. Oz (2015) explores pre-service English teachers' TPACK through a TPACK scale (Schmidt et al., 2009) with open-ended questions. The findings highlight that the participants develop their TPACK significantly. Similarly, Kwangsawad (2016) investigates pre-service English teachers' TPACK through a TPACK survey (Schmidt et al., 2009), lesson plans, and classroom observations in Thailand. The research shows that the participants have high scores in all domains of TPACK. Additionally, in a qualitative case study, Wetzel and Marshall (2011) explore in-service English teachers' performances on TPACK. The data for the research is collected through classroom observations and interviews. The study concludes that the teacher can display classroom management practices well. Wu and Wang (2015) examine TPACK of in-service English teachers through self-reported questionnaire, interviews and classroom observations. The results indicate that EFL teachers are confident in their PK and they need more technological knowledge to further develop their TPACK level. In a mixed-method study, Liu and Kleinsasser (2015) question in-service English teachers' TPACK and perceived computer self-efficacy in CALL training courses. In the study, a survey, interviews, and posted messages are used as data collection instruments. Data analysis shows an increase in in-service English teachers' TPK, TCK, TPACK ratings and computer self-efficacy scores. Rubadeau (2016) analyses cognitions and practices on the integration of pedagogy and technology of in-service English teachers. Data collection process is carried out through semi-structure interviews, classroom observations, written reflections, field notes, and documents reviews. The findings of the study emphasise that the participants show high levels of TPACK. Also, in a longitudinal study questioning whether pre-service teachers' perceived increase in TPACK skills follows a linear increase in four-year-long language education program, data for the study is collected through a TPACK survey with open-ended questions. The results of the study underline that there is a nonlinear pattern of TPACK development in four-year-long education process (Turgut, 2017a).

Literature review reveals that a number of researchers have made an attempt to measure perceptions, self-efficacy, competency, and skills of pre-service and in-service English teachers through various data collection instruments including self-reporting surveys/questionnaires, open-ended questionnaires, interviews, and observations based on the framework of Teachers' Knowledge of Teaching and Technology (TKTT), which is frequently employed as the key instrument designed by Mishra and a group of researchers (Schmidt et al., 2009; Young et al., 2013)

Apart from its contribution to serving as an instrument to measure knowledge of English teachers, TPACK can also play an important role in revealing required competencies/skills to develop curricula in line with TPACK dimensions for pre-service English teachers and design professional development trainings for in-service English teachers in the 21st century. Using reliable and valid TPACK instruments as a lens for evaluating English teachers' knowledge may also have effect on quality of language teaching and design of professional development. Hence, in order to provide more accurate insights into the way how to better equip pre-service and in-service English teachers with required competencies/skills based on TPACK in the 21st century, it is essential to investigate how researchers in the field of EFL address the issues of reliability and validity of TPACK instruments in their studies. In addition to this, since there is the paucity of studies questioning how researchers in the field of EFL address the reliability and validity of TPACK instruments, to fill the gap in this field, the researcher intends to seek the evidence of reliability and validity of instruments reported in each of the reviewed studies through the following research questions:

(1) What instruments are employed to measure TPACK in the reviewed studies?

(2) Are the instruments reliable and valid to measure TPACK in the reviewed studies?

2. METHOD

2.1. Search Strategies and Procedure

To seek answers for the research questions, the search was performed on ERIC, ScienceDirect, Scopus, EBSCOhost, and Web of Science databases. Each search was repeated on the databases to check possible selection bias and then a comparison of the obtained studies was made. Afterwards, studies were identified where (a) TPACK was discussed in terms of pre-and in-service English teachers through titles, keywords, or abstracts. In order to obtain comprehensive search results, the keywords for each search were "technological pedagogical content knowledge", "TPCK" "technological pedagogical and content knowledge", and "TPACK" The search was limited to studies published between 2010 and 2019 in order to cover as many studies as possible.

2.1.1. *Inclusion Criteria*

A set of inclusion and exclusion criteria was employed in the process of publication selection (Table 1). Articles, full-text conference papers, and dissertations written in English were included. Other types of studies such as editorials, theoretical studies/reviews, book chapters, and other studies irrelevant to the focus of this review were excluded. The initial search yielded 235 studies. Firstly, the abstracts of the 235 studies were read and reviewed by the researcher. In case of any ambiguity, the study was completely read. After the inclusion and exclusion criteria were applied to yielded studies in line with the research questions, a quite few theoretical studies/reviews were excluded since they were irrelevant to the focus of this study. In addition to this, studies discussing TPACK from different perspectives were left out. As a result of the initial review of 235 studies, 75 studies remained for the researcher to complete reading of them in-depth. In the event of borderline, an external researcher with insight into this field was also consulted to read the study. From the full-text reading, 60 studies (including 40 articles, 14 dissertations and 6 conference papers) were chosen for thorough analysis.

Table 1. *Inclusion and exclusion criteria*

Inclusion	Exclusion
Articles	Studies available in summary
Full-text conference papers	Editorials and summary reports
Dissertations	Book chapters
Studies in EFL with TPACK instruments	Theoretical studies / reviews

2.1.2. *Data Coding Scheme*

A total of 60 studies were applied to content analysis. The overall characteristics of the publications are classified according to a set of criteria including publication year, types of publication, instrument types, research design, reliability, validity, and target audience. At the study level, publication year, types of study, instrument types, and types of research design in each study were listed (Table 2 & 3). At the instrument-level analysis, target audience, reliability, and validity of each TPACK instrument were checked and testing process of each instrument's reliability and the validity was then reported.

Coding process was carried out by the researcher. When there was an ambiguous case, an external researcher with insight into content analysis and coding was consulted. A total of 60 studies were included for coding process. To establish the robustness of the coding, randomly selected 15 studies were coded independently by an external researcher. As a result of separate coding process, a high agreement (inter-coder reliability .89) was reached by the researchers. As the majority of the studies employed more than one type of TPACK instrument, each study in the review process was coded multiple times. For instance, in the research conducted by Abera (2014), interviews, classroom observations, documents, and a questionnaire were employed to reveal TPACK level of English teachers at tertiary level. For this reason, the study was coded four times since there were four different instruments in the same study.

3. FINDINGS

3.1. Study-level analysis

Most of the reviewed studies are articles and full-text conference papers (46 out of 60 studies). The remainder of the studies (14) consists of unpublished doctoral and master dissertations. The number of the studies into the use of TPACK in the field of ELT increases each year (Table 2). As for the kinds of TPACK instruments, more than two different types of TPACK instruments are identified in nearly half of the studies (a total of 128 out of 60).

Classification of studies in terms of research designs in the reviewed studies shows that half of the articles (20), all the conference papers (6), and half of the dissertations (7) are conducted by quantitative research designs in the reviewed studies. In seventeen of the articles and five of the dissertations, the researchers carry out their research based on qualitative research designs. For the remainder of articles (3) and dissertations (2), the researchers report mixed methods research design in their studies. Considering the types of studies in terms of research design, it is revealed that the researchers generally prefer to design their research based on quantitative and qualitative research designs rather than mixed method research designs.

To find out how each researcher addresses reliability and validity issues of each TPACK instrument and provides evidence of reliability and validity in their studies, a two-level analysis is conducted. First level analysis is based on revealing general characteristics (including types of study, publication years, types of TPACK instruments, and types of research design) of each study in order to have a complete understanding about their studies (Table 2 & 3). At the instrument-level analysis, together with the target audience each TPACK instrument is examined in terms of reliability and validity (Table 4).

Table 2. *Characteristics of the (N=60) studies in the review.*

Category	Number	%
Study type		
Article	40	67%
Conference paper	6	10%
Dissertation	14	23%
Publication Year		
2011	1	2%
2012	2	4%
2013	5	8%
2014	9	15%
2015	11	18%
2016	8	13%
2017	11	18%
2018	7	12%
2019	6	10%
Instruments		
1	15	25%
2	18	30%
3	19	32%
4	5	8%
5	3	5%

Table 3. *Classification of studies in terms of research designs in the review*

Study type	Quantitative	Qualitative	Mixed method
Articles	N=20	N=17	N=3
Conference papers	N=6	N=0	N=0
Dissertations	N=7	N=5	N=2

3.2. Instrument-Level Analysis

Following study-level analysis, each TPACK instrument is counted in the reviewed studies. It is seen that there are five types of instruments that are not evenly distributed in the reviewed studies. Self-report instruments (60), interviews (32), and observations (21) are reported to be most used ones, whereas open-ended questionnaires (9) are identified to be the least preferred TPACK instruments in the reviewed studies (Table 4).

Table 4. *The description of instruments in terms of target audience, reliability, and validity*

Instruments	Self-report	Open-ended	Performance	Interview	Observation
Number of instruments	N=60, 46%	N=9, 7%	N=14, 11%	N=25, 20%	N=20, 16%
Target audience					
Pre-service	N=22, 37%	N=4, 44%	N=5, 36%	N=6, 24%	N=5, 25%
In-service	N=34, 57%	N=4, 44&	N=9, 64%	N=18, 72%	N=12, 60%
Pre & in service	N=4, 6%	N=1, 12%	N=0, 0%	N=1, 4%	N=3, 15%
Reliability					
Clearly presented	N=40, 67%	N=4, 44%	N=3, 21%	N=2, 8%	N=2, 10%
Not presented	N=20, 33%	N=5, 56%	N=11, 79%	N=23, 92%	N=18, 90%
Validity					
Clearly presented	N=24, 40%	N=0, 0%	N=0, 0%	N=0, 0%	N=0, 0%
Not presented	N=36, 60%	N=0, 0%	N=0, 0%	N=0, 0%	N=0, 0%

3.2.1. *Self-Report Instruments*

Self-report instruments like Thurstone scales or Likert scales are regarded as the instruments in which participants are required to report directly on their own behaviours, beliefs, attitudes, or intentions (Lavrakas, 2008). As well, as the source of obtaining quantitative research data, self-report instruments like surveys or questionnaires should be proven to be valid, reliable, and unambiguous in the process of designing (Richards & Schmidt, 2002).

Nearly half of the instruments (60) are self-reported instruments that are used to assess TPACK of English teachers. More than half of the self-report instruments aim to measure TPACK of in-service English teachers. The four of the self-reported instruments are employed for the purpose of assessing both pre-and in-service English teachers (Drajati et al., 2018; Tseng et al., 2019; Turgut, 2017b; Wang, 2016). Most of the self-report instruments cover multiple sub-scales of TPACK framework. To illustrate, Vereshchahina et al. (2018) employ TPACK survey to analyse self-assessment of English instructors. The self-report TPACK instrument is composed of 39 items and 7 sub-scales based on TPACK framework. The study questions whether English teachers can successfully combine the content of English language and language teaching methods with sufficient use of computer technologies in order to achieve educational goals.

Forty of the studies provide the index of reliability based on cronbach's alpha. For example, Kharade and Peese (2014) express the reliability of the seven domains ranging from .83 to .93. As for validity, in less than half of the self-report instruments (24 out of 60) validity is established mostly through either exploratory or confirmatory factor analysis. For instance, in order to test of validity of TPACK-EFL, which is regarded as an assessment tool for teachers of English as foreign language (EFL), firstly survey items are constructed through mixed methods research design. The process of content validity of the items is conducted through expert and pre-service teacher reviews and then to validate the survey two rounds of exploratory factor analysis are carried out. The first-round analysis shows that the survey is composed of five-factor structure: technological knowledge (TK), content knowledge (CK), pedagogical knowledge (PK), pedagogical content knowledge (PCK). There is also the fifth factor combining TCK, TPK, and TPACK items. Upon making revisions on the survey, the second round of analysis shows that there is a seven-factor structure consistent with the framework of

TPACK. The TPACK-EFL survey includes 39 items in total. Under the dimension of TK, CK, PK, PCK, TCK, TPK, and TPACK, there are 9, 5, 6, 5, 3, and 4 items respectively (Baser et al., 2016).

3.2.2. *Open-Ended Questionnaires*

Open-ended questionnaires refer to a set of questions whose responses/answers are constructed by interviewees (Lewis-Beck et al., 2004). Such questions can lead to a greater level of valuable discovery of information from the perspectives of respondents in qualitative research designs; however, since their open-ended nature makes it difficult to reflect what respondents mean to say, the issue of reliability and validity is of vital importance to researchers in order to yield as accurate and reliable data as possible (Nunan, 1999).

In the reviewed studies, only nine (out of 128) open-ended questionnaires are identified. In addition to this, only one of the open-ended questionnaires targets pre-service and in-service English teachers (Turgut, 2017b). In her study, the open-ended questionnaire is employed to investigate the participants' perceptions of how TPACK is modelled by English teachers. The aim of the open-ended questionnaire with three questions is to examine whether English teachers effectively display the integration of content, technology with teaching methods in the classroom. In the light of the responses of the participants, codes and themes are created by the researcher. Only four out of nine open-ended questionnaire instruments express inter-rater reliability as evidence of reliability. However, the issue of validity is not explicitly addressed in any of the open-ended questionnaire (Table 4).

3.2.3. *Performance Assessments*

Performance assessment describes an approach which requires participants to construct or perform an original response in accord with given authentic tasks or realistic scenarios (Frey, 2013; Good, 2008). Only 18 out of 128 instruments in the reviewed studies are identified as performance assessments. All of the instruments of performance assessments are designed to evaluate either pre-service or in-service English teachers. In some of TPACK performance assessment tasks, English teachers are asked to prepare a set of artefacts like teaching syllabi, instructional materials and reflective journals aiming to investigate the effectiveness and quality of English teachers' implementation of their teaching in line with the framework of TPACK (Alhababi, 2017), whereas in other TPACK performance assessment tasks English teachers are required to create a set of teaching artefacts such as web portfolios and digital stories to evaluate the effectiveness of TPACK framework (Harriman, 2011) and teachers' digital literacies (Weerakanto, 2019). In the reviewed studies, only three instruments of performance assessment present evidence of reliability through the inter-rater reliability (Chewning, 2015; Ersanli, 2016; Le & Song, 2018). None of the instruments of performance assessment provide any evidence of validity.

3.2.4. *Interviews*

An interview is a situation where the interviewer asks the interviewee a set of questions that are generally done face-to-face or over the telephone or recorded in audiotapes or videotapes for transcription. In addition, interviews are sometimes possible to be electronically conducted, such as over the Internet (Johnson et al., 2014; Johnson & Christensen, 2019; Gall et al., 2007). Considered to be one of the most frequently used instruments for qualitative data collection, an interview is a valuable method for questioning people's views and their meanings in a natural setting (Cohen et al., 2007). As Dörnyei (2007) avers, validity and reliability issues of these instruments serve as guarantees of research results and accuracy of data.

In total, 14 interview instruments (out of 25) do not provide any explicit and detailed information. Nine of the interviews are conducted in a semi-structured way. Only two interview types (out of 25) are performed through a focus group interview where a group moderator guides a talk with a group of people such as students, or teachers to make them discuss the topic. The moderator also forms group talks with the help of open-ended questions by acting as a facilitator of the group (Johnson et al., 2014; Johnson & Christensen, 2019; Gall et al., 2007). For example, Asık et al. (2018) aim to get a detailed understanding of pre-service English teachers' use of digital tools, and each of the researchers conducts three focus group interviews with a total of 30 randomly selected participants. In focus group interviews, the participants are asked six questions prepared by the researchers in advance. With the permission of the participants, the researchers make the record of the interviews and then the record is prepared for analysis. In the reviewed studies, only one researcher (Alahmari, 2013), who questions English teachers' use of technology, their willingness to use technology, and their perceptions of TPACK, conducts the interviews electronically over Skype with 10 participants about 20 minutes on average. As for reliability of the instruments, only two of the studies out of 25 reports concrete evidence of reliability based on inter-rater reliability. In both of the studies, the percent agreement for two coders is .77 (Ansyari, 2012, 2015). Additionally, none of the studies provide an explicit evidence of validity (Table 4).

3.2.5. *Observations*

An observation means watching relevant phenomena by taking extensive field notes in both qualitative and quantitative research paradigms. Researchers record what is believed to be important in their field notes. In observational activities in the field, videotaping or audiotaping could also be employed to record necessary parts of observations (Johnson et al., 2014; Johnson & Christensen, 2019; Gall et al., 2007). While using less structured observation instruments in qualitative research designs, accuracy and consistency of observational data might be a threat to researchers who attempt to ensure good reliability and validity for their research results.

In the reviewed studies, only two out of 20 studies report the use of video recording (Kharade & Peese, 2014; Weerakanto, 2019). In one of those studies, the researcher intends to identify the perceptions of pre-service English teachers and the researcher also examines how the teachers apply technology to their pedagogical practices. Hence, the researcher conducts two class observations by video recording the teachers in three English language classrooms during nine weeks as a non-participant observer (Weerakanto, 2019). The video record is then transcribed to be examined and coded by the researcher. In 18 studies out of 20, the researchers take field notes during their observations in order to shed crucial light on how English teachers apply their knowledge of pedagogy, content, and technology in their classroom settings.

For instance, in the study of Tai (2015) in order to both understand how English teachers integrate technology into classroom teaching and identify how classroom activities are appropriately integrated with pedagogical approaches, the researcher employs an observation instrument including three sections: (1) Background Part, which gives a brief information regarding the role of the observer in the context and content (2) Competency Part including TPACK items for directing observations, and (3) Post Observation Part, which is for taking notes and writing down questions during observations. A total of 26 classes of thirteen English teachers are observed and then observation field notes are sorted into units of analysis to be examined and coded by the researcher.In the reviewed studies, only two of the studies using observation instruments perform the reliability of the instruments (Chewning, 2015; Tai, 2015) through the index of inter-rater reliability and report inter-coder reliabilities as .81 and .78 respectively. For validity, none of the studies provide any explicit evidence of validity (Table 4).

4. DISCUSSION and CONCLUSION

It is revealed that out of 128 instruments in 51 instruments, the reliability of those instruments is ensured through Cronbach's alpha and inter-rater reliability. Besides, the validity of the instruments is performed through expert content validity and factor analyses. Given that the number of studies based on quantitative research design in the reviewed studies, it is unsurprising to find out that the distribution of self-report instruments is nearly half of (46%) the total instruments. In a quantitative research design, survey research employs some sort of surveys or questionnaires to describe attitudes, opinions, perceptions or experiences (Creswell, 2005; Mertens, 2005). The majority of the reviewed studies underlines that the researchers utilise self-report instruments designed based on TPACK framework to investigate pre- and in-service English teachers' perceptions, beliefs, and self-efficacy. As Mertens (2005) explains, self-report instruments are used as the descriptive surveys to describe the characteristics of a group at one point in time.

The crucial point concerning the collected data through self-report instruments is that a self-report instrument by its very nature makes researchers trust what participants believe is true or what they have experienced. In view of Leedy and Ormrod (2013), researchers need to remember two important issues – reliability and validity when it comes to collecting self-reported data. Similarly, Winter (2000) also states that reliability and validity are tools of an essentially positivist epistemology. Thus, it might be more appropriate for researchers to select positivist research for their research since positivism, to some extent, is defined by a systematic theory of validity (Joppe, 2000), through which researchers truly measure what they intend to measure and ensure truthful outcomes regarding TPACK level of English teachers. Whereas reliability and validity are the terms of positivist quantitative paradigm that refer to the replicability and accuracy of measures, credibility and trustworthiness are the constructs of qualitative paradigm (Merriam & Tisdell, 2015; Saldana, 2011). That is to say, qualitative research is based on assumptions of a researcher about reality different from those of quantitative research. Taken the novelty of TPACK in the field of EFL and intricate nature of TPACK framework into consideration, it would not be a viable solution for researchers to employ solely qualitative paradigms in their research.

The employment of interview as an instrument to gather research data is in the second place, which shows that its use is slightly higher than that of observation instrument (Table 4). In the reviewed studies, interview data is collected through focus-group interviews and semi-structured interviews. Considering the challenges of data analysis of interviews, it is not surprising to find out that a very limited number of studies report reliability and none of those studies ensure the validity. Albeit interview's elusive nature as an instrument (Creswell, 2009), in order to increase its reliability and validity in qualitative studies, a try-out of the interview protocol, which is also known as a trial run is expected to be conducted by researchers prior to a full-scale study (Teijlingen van & Hundley, 2001). In every research design, instruments chosen for data collection are supposed to pass the tests of validity and reliability before they can be considered to be good measures, hence the conduct of a pilot study as fundamental to any research needs to be crucial for researchers in the field of English language teaching. A pilot interviewing may enable researchers to identify ambiguities with unnecessary questions, specify if each question elicits a sufficient response (Teijlingen van & Hundley, 2001), and most importantly allow researchers to practise and perfect interviewing techniques prior to real research settings (Berg, 2001).

As for observation as a data collection instrument in the reviewed studies, both quantitative and qualitative observations are employed by the researchers; however, only two of the studies report the reliability of observation instruments with no proof of validity provided by the researchers. The researchers conducting quantitative observation employ checklists and videotape recorders to record data for coding later. As well, some of the researchers in the reviewed studies utilise a naturalistic observation in classroom settings where they take on the role of observer much more than a participant (Johnson & Christensen, 2019). Good reliability in an observation protocol depends on the consistency of observations across time and observers. Likewise, good validity in an observation protocol ensures that observation instrument measures what it is intended to measure (Maxwell, 2012). In other words, the reliability of an observed behaviour is also closely linked to the validity of the observation. Gardner (2000) asserts that reliability of an instrument imposes limits on its validity. To put it another way, lack of a valid protocol for observation especially in qualitative research design makes the reliability of the instrument ineffective (DeMonbrun et al., 2015).

Open-ended questionnaires are the least employed instrument in the reviewed studies. An instrument of open-ended questionnaire can prompt a lengthy and detailed response, much of which could not be relevant to the topic and might be hard to code for a researcher (Lewis-Beck et al., 2004). Similarly, in view of Koehler et al. (2012) the difficulties of coding and analysing data of open-ended questionnaire instruments could be among the important reasons why it is the least preferred instrument by the researchers.

The second least preferred instrument is performance assessment instruments. Performance assessment includes teaching syllabi, instructional materials, and reflective journals that are employed to identify how much the participants could put TPACK into practice in their acts of teaching. In particular the use of reflective journals in teacher education enables the researcher to make strong relationship with the participants (O’Connell & Dyment, 2011) by providing the researcher with an opportunity to hear the voice of them through their reflections while gaining practical TPACK experiences (Dunlap, 2006). As a means of data collection instrument in qualitative research designs, a reflective journal may also enable the researcher to evaluate the contributions of TPACK-related training or practices to English teachers. To the best of my knowledge, the challenge for the researcher lies in the difficulty of analysing and coding qualitative data gathered through reflective journals.

Considering the numbers of data collection instruments in the reviewed studies, qualitative data collection instruments in total are more than quantitative data collection instruments; however, self-report instruments provide higher ratio of reliability and validity when compared with that of all qualitative data collection instruments in the study. These issues in quantitative research design are dependent upon the construction of an instrument; however, in qualitative research design, the researcher is the instrument (Patton, 2001). Moreover, in qualitative research design what is largely missing in the literature for researchers is certainty about whether they are supposed to make an agreement based on codes, themes, or both codes and themes (Creswell & Poth, 2016). This may also account for less employment of qualitative research instruments than self-report instruments in the reviewed studies.

Given TPACK framework, it provides a theoretical background for teacher education that aims to integrate good teaching with technology by integrating technological, pedagogical, and content knowledge (Koehler & Mishra, 2005). However, not being thoroughly cognisant of what TPACK framework offers owing to its complex and overlapping structure, most of the researchers use TPACK as self-report instrument in their research to measure participants’ perceptions, self-efficacy, competencies, and skills. Researchers in the field of English language teaching are expected to design and develop quantitative or qualitative data collection instruments that help measure how much English teachers truly demonstrate teaching activities, performances, and professional learning.

Another point that could be raised why the researchers employ mostly self-report instruments instead of other instruments is that TPACK is a complicated framework and covers multiple domains. According to Koehler et al. (2012), as TPACK is composed of multiple domains and intersections, it requires sophisticated understanding of the domains and intersections for researchers to customise TPACK to a specific field of research and devise any kind of instrument. In the same vein, a group of scholars (Chai et al., 2010; Cox & Graham, 2009) find it difficult to pinpoint the distinction of each of the domains (PCK, TCK, and TPK) as the boundaries between them are quite fuzzy. Hence, complexities of distinguishing between those domains might make the development of a valid and reliable instrument also difficult for researchers in this field. In addition, another issue concerning why reliability and validity of the instruments occur in the reviewed studies is that the use of TPACK in this field has just started to emerge (Le & Song, 2018; Öz, 2015). This might be another explanation for inadequate number of instruments with the evidence of reliability and validity.

As an alternative to ensuring reliability and validity of instruments, triangulation seems to be a solution; however, according to Seawright (2016), triangulation in social sciences has considerable flaws. In the current study, for example in order to measure TPACK of English teachers the researchers collect data based on qualitative and quantitative research designs through different instruments including different questions even though they concentrate on the same TPACK framework. The use of instruments with different questions makes both the reliability and validity of the instruments and research findings problematic since the employment of quantitative and qualitative instruments including different questions may generate different findings. In his view (Seawright, 2016), the focal point of integrative multi-method research is to utilise each research method for what it is especially good at and to minimise inferential weaknesses by using other methods to test, revise, or justify assumptions. Thus, integrative designs employing multiple modes of inference to substitute strengths for weaknesses could be another solution especially for researchers who may have difficulty in ensuring reliability and validity of instruments in this field.

To sum up, though self-report instruments are highly versatile and relatively easy to employ, one of the weaknesses of self-report instruments is that participants may have an inclination to express themselves more differently than they really are (Bordens & Abbott, 2011). In qualitative studies the researchers are required to follow rigorous data collection and challenging data analysis processes based on their assumptions that influence quality and the results of the research (Gibbs et al., 2007; Kitto et al., 2008). Therefore, it might be supposed by the researchers that utilising quantitative and qualitative data collection instruments together in their studies would naturally resolve the issues of reliability and validity of such instruments as interviews and open-ended questionnaires. The reasons why a limited number of instruments like interview and open-ended questionnaire ensure reliability and validity might be attributed to meticulous data collection and challenging data analysis processes in qualitative research design or the researchers' assumption of triangulation. Besides, the complexities of measuring performance and real-life scenario tasks might prompt the researchers to use other instruments instead of performance assessment instruments.

Finally, since TPACK is newly emergent scope of research for researchers in the field of EFL, some issues like ensuring reliability and validity of instruments in either quantitative or qualitative research designs could appear to be exhausting and challenging, thus researchers could welcome integrative multi-method research designs as a panacea for especially minimising reliability and validity issues of their instruments and producing more reliable and accurate research results.

Despite the fact that TPACK has come under widespread criticism from scholars and researchers in every field of research, it is an undeniable fact that TPACK has made substantial contributions to the field of education by presenting a framework to question teachers' knowledge of content, pedagogy, and technology. Also, to the best of my knowledge, TPACK offers an opportunity for teachers to replace traditional teaching methods with technology integrated ones to be able to perform their professions more efficiently and more effectively.

5. RECOMMENDATIONS

Future researchers should develop new TPACK instruments capable of measuring actual learning, performance, and real-life scenario tasks apart from the ones used to measure perception, belief self-efficacy through TPACK instruments. In addition to the use of Cronbach's alpha, inter-rater reliability, expert content validity, and factor analysis to ensure the reliability and validity of instruments, future researchers should also try using other ways of ensuring and increasing reliability and validity of instruments while devising new TPACK instruments to measure TPACK of English teachers. Future researchers should meticulously look into the ways how multi-method research designs and mixed methods research designs could be employed to measure English teachers' TPACK in further studies. Further researchers should also question how data triangulation process in a TPACK-research works in terms of reliability and validity and might be applied to better measure pre-service and in-service English teachers' TPACK.

Limited to investigate reliability and validity issues, this review has made an attempt to discuss how the issues of reliability and validity of instruments are addressed by the researchers within a limited number of studies in the field of EFL.

Declaration of Conflicting Interests and Ethics

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

ORCID

Abdullah ARSLAN https://orcid.org/0000-0002-3979-6371

6. REFERENCES

Abera, B. (2014). Applying a technological pedagogical content knowledge framework in Ethiopian English language teacher education. *In Multicultural Awareness and Technology in Higher Education: Global Perspectives*, 286-301. IGI Global.

Agyei, D. D., &Voogt, J. (2012). Developing technological pedagogical content knowledge in pre-service mathematics teachers through collaborative design. *Australasian journal of educational technology*, 28(4), 547-564.

Alahmari, A. S. (2013). *An investigation of Saudi Arabian EFL teachers' engagement with technology*. (Unpublished doctoral dissertation). Retrieved from https://monash.figshare.com/ 4701100_monash_120645.pdf

Alhababi, H. H. (2017). *Technological pedagogical content knowledge (TPACK) effectiveness on English teachers and students in Saudi Arabia*. (Unpublished doctoral dissertation). Retrieved from https://digscholarship.unco.edu/dissertations/456/

Ansyari, M. H. (2012). *The development and evaluation of a professional development arrangement for technology integration to enhance communicative approach in English language teaching* (unpublished master's thesis). Retrieved from https://essay.utwente.nl/62294/1/MSc_Ansyari_M.F._-_S1084712.pdf

Ansyari, M. F. (2015). Designing and evaluating a professional development programme for basic technology integration in English as a foreign language (EFL) classrooms. *Australasian Journal of Educational Technology*, *31*(6), 699-712.

Archambault, L., & Crippen, K. (2009). Examining TPACK among K-12 online distance educators in the United States. *Contemporary issues in technology and teacher education, 9*(1), 71-88.

Aşık, A., Eroğlu İnce, B., & Şarlanoğlu Vural, A. (2018). Investigating learning technology by design approach in pre-service language teacher education: Collaborative and reflective experiences. *Eğitimde Nitel Araştırmalar Dergisi,6*(1), 37-53.

Baser, D., Kopcha, T. J., &Ozden, M. Y. (2016). Developing a technological pedagogical content knowledge (TPACK) assessment for preservice teachers learning to teach English as a foreign language. *Computer Assisted Language Learning, 29*, 749–764.

Berg, B. L. (2001). *Qualitative research methods for the social sciences* (4th ed.). Boston, MA: Allyn and Bacon.

Bordens, K. S., & Abbott, B. B. (2002). *Research design and methods: A process approach*. McGraw-Hill.

Bowers, J. S., & Stephens, B. (2011). Using technology to explore mathematical relationships: A framework for orienting mathematics courses for prospective teachers. *Journal of Mathematics Teacher Education, 14*(4), 285-304.

Canbazoglu-Bilici, S., Yamak, H., Kavak, N., & Guzey, S.S. (2013). Technological pedagogical content knowledge self-efficacy scale (TPACK-SeS) for pre-service science teachers: construction, validation, and reliability. *Eurasian Journal of Educational Research, 52*, 37– 60.

Chai, C. S., Koh, J. H. L., & Tsai, C.-C. (2010). Facilitating preservice teachers' development of technological, pedagogical, and content knowledge (TPACK). *Educational Technology & Society, 13*(4), 63–73.

Chai, C. S., Chin, C. K., Koh, J. H. L., & Tan, C. L. (2013). Exploring Singaporean Chinese language teachers’ technological pedagogical content knowledge and its relationship to the teachers’ pedagogical beliefs. *The Asia-Pacific Education Researcher, 22*(4), 657-666.

Chai, C. S., Koh, J. H. L., & Tsai, C. C. (2013). A review of technological pedagogical content knowledge. *Journal of Educational Technology & Society, 16*(2), 31-51.

Chewning, R. (2015). *Secondary English teachers’ dispositions toward technology integration in one-to-one environments* (Unpublished Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global database. (UMI No. 3745350)

Cohen, L. (2007). Experiments, quasi-experiments, single-case research and meta-analysis (Cohen, L., Manion, L., & Morrison, K. in Eds) *Research methods in education*. (6th ed.).

Cox, S., & Graham, C. R. (2009). Using an elaborated model of the TPACK framework to analyse and depict teacher knowledge. *TechTrends, 53*(5), 60-69.

Creswell, J. W., &Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Creswell, J. W. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Pearson Education, Inc.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Sage Publications.

DeMonbrun, M. R. M., Finelli, C. J., & Shekhar, P. (2015). Methods for establishing validity and reliability of observation protocols. In *ASEE Annual Conference and Exposition, Conference Proceedings* (Vol. 122, No. 122, pp. 1-10).

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics: Quantitative Qualitative, and Mixed Methodologies*. Oxford: Oxford University Press.

Drajati, N. A., Tan, L., Haryati, S., Rochsantiningsih, D., &Zainnuri, H. (2018). Investigating English language teachers in developing TPACK and multimodal literacy. *Indonesian Journal of Applied Linguistics, 7*(3), 575-582.

Dunlap, J. C. (2006). Using guided reflective journaling activities to capture students' changing perceptions. *Techtrends: Linking Research & Practice to Improve Learning,* 50, 20–26. https://doi.org/10.1007/s11528-006-7614-x

Ersanli, C. Y. (2016). Improving technological pedagogical content knowledge (TPACK) of pre-service English language teachers. *International Education Studies, 9*(5), 18-27.

Frey, B. B. (2013). *Modern classroom assessment*. Sage publications.

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational Research: An Introduction*. New York: Person Education.

Gardner, F. (2000). Methodological issues in the direct observation of parent–child interaction: Do observational findings reflect the natural behavior of participants?. *Clinical child and family psychology review, 3*(3), 185-198.

Gibbs L., Kealy M., Willis K., Green J., Welch N. & Daly J. (2007) What have sampling and data collection got to do with good qualitative research? *Australian and New Zealand Journal of Public Health, 31*(6), 540-544. https://doi.org/10.1111/j.1753-6405.2007.00140.x

Graham, C. R. (2011). Theoretical considerations for understanding technological pedagogical content knowledge (TPACK). *Computers & Education, 57*(3), 1953-1960.

Graham, C. R., Borup, J., & Smith, N. B. (2012). Using TPACK as a framework to understand teacher candidates' technology integration decisions. *Journal of Computer Assisted Learning, 28*(6), 530-546.

Good, T. L. (Ed.). (2008). *21st-century education: A reference handbook*. Vol.1. Sage.

Harriman, C. L. S. (2011). *The impact of TPACK and digital storytelling as a learning experience for pre-service teachers in a learning-by-designing project* (unpublished doctoral dissertation). University of Georgia, USA.

Harris, J., Grandgenett, N., & Hofer, M. (2010, March). Testing a TPACK-based technology integration assessment rubric. *In Society for Information Technology & Teacher Education International Conference*, 3833-3840. Association for the Advancement of Computing in Education (AACE).

Jang, S. J., & Tsai, M. F. (2012). Exploring the TPACK of Taiwanese elementary mathematics and science teachers with respect to use of interactive whiteboards. *Computers & Education, 59*(2), 327–338.

Jen, T. H., Yeh, Y. F., Hsu, Y. S., Wu, H. K., & Chen, K. M. (2016). Science teachers' TPACK-Practical: Standard-setting using an evidence-based approach. *Computers & Education, 95*, 45–62.

Johnson, R. B., & Christensen, L. (2019). *Educational research: Quantitative, qualitative, and mixed approaches*. SAGE publications.

Johnson, R. B., Christensen, L. B., & Turner, L. A. (2014). *Research methods, design and analysis*. Pearson Education.

Joppe, M. (2000). The Research Process: Tests and Questionnaires. *Quantitative Applications in the Social Sciences*, 211-236.

Kafyulilo, A., Fisser, P., Pieters, J., &Voogt, J. (2015). ICT use in science and mathematics teacher education in Tanzania: Developing technological pedagogical content knowledge. *Australasian Journal of Educational Technology, 31*(4), 381–394.

Kitto S. C., Chesters J. & Grbich C. (2008) Quality in qualitative research. *Medical Journal of Australia, 188*(4), 243-246.

Kramarski, B., & Michalsky, T. (2010). Preparing preservice teachers for self-regulated learning in the context of technological pedagogical content knowledge. *Learning and Instruction, 20*(5), 434-447.

Kharade, K., &Peese, H. (2014). Problem-based learning: A promising pathway for empowering pre-service teachers for ICT-mediated language teaching. *Policy Futures in Education, 12*(2), 262-272.

Koçoğlu, Z. (2009). Exploring the technological pedagogical content knowledge of pre-service teachers in language education. *Procedia-Social and Behavioural Sciences*, *1*(1), 2734-2737.

Koehler, M. J., & Mishra, P. (2008). Introducing TPCK. AACTE Committee on Innovation and Technology (Ed.), *The handbook of technological pedagogical content knowledge (TPCK) for educators*, 3-29. Lawrence Erlbaum Associates.

Koehler, M., & Mishra, P. (2009). What is technological pedagogical content knowledge (TPACK)? *Contemporary issues in technology and teacher education, 9*(1), 60-70.

Koehler, M. J., Shin, T. S., & Mishra, P. (2012). How do we measure TPACK? Let me count the ways. *In educational technology, teacher knowledge, and classroom impact: A research handbook on frameworks and approaches*, 16-31. IGI Global.

Koh, J. H. L., Chai, C. S., & Tsai, C. C. (2013). Examining practising teachers' perceptions of technological pedagogical content knowledge (TPACK) pathways: A structural equation modelling approach. *Instructional Science, 41*(4), 793-809.

Kurt, G., Akyel, A., Koçoğlu, Z., & Mishra, P. (2014). TPACK in practice: A qualitative study on technology integrated lesson planning and implementation of Turkish pre-service teachers of English. *ELT Research Journal*, *3*(3), 153-166.

Kwangsawad, T. (2016). Examining EFL Pre-service Teachers' TPACK Trough Self-report, Lesson Plans and Actual Practice. *Journal of Education and Learning*, *10*(2), 103-108.

Landry, G. A. (2010). *Creating and validating an instrument to measure middle school mathematics teachers' technological pedagogical content knowledge. (TPACK)*. (Unpublished doctoral dissertation). University of Tennessee. Retrieved from http://trace.tennessee.edu/utk_graddiss/720.

Lavrakas, P. J. (2008). *Encyclopaedia of survey research methods*. Sage publications.

Le, N., & Song, J. (2018). TPACK in a CALL course and its effect on Vietnamese pre-service EFL teachers. *The Asian EFL Journal Quarterly*, 31.

Lewis-Beck, M., Bryman, A. E., & Liao, T. F. (2004). *The Sage encyclopaedia of social science research methods*. Sage Publications.

Liu, M. H., & Kleinsasser, R. (2015). Exploring EFL teachers' knowledge and competencies: In-service program perspectives. *Language Learning & Technology*, *19*(1), 119-138.

Luppicini, R. (2005). A systems definition of educational technology in society. *Educational Technology & Society, 8*(3), 103-109.

Maxwell, J. A. (2012). *Qualitative research design: An interactive approach*. Sage.

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers' college record, 108*(6), 1017-1054.

Mouza, C., Karchmer-Klein, R., Nandakumar, R., Ozden, S. Y., & Hu, L. K. (2014). Investigating the impact of an integrated approach to the development of preservice teachers' technological pedagogical content knowledge (TPACK). *Computers & Education, 71*, 206–221.

Niess, M. L. (2011). Investigating TPACK: Knowledge growth in teaching with technology. *Journal of educational computing research, 44*(3), 299-317.

Nunan, D. (1999). *Research methods in language learning* (8th printing). Cambridge: CUP.

O'Connell, T. S., &Dyment, J. E. (2011). The case of reflective journals: Is the jury still out? *Reflective Practice*, 12, 47–59. https://doi.org/10.1080/14623943.2011.541093

Oz, H. (2015). Assessing pre-service English as a foreign language teachers' technological pedagogical content knowledge. *International Education Studies, 8*(5), 119-130.

Patton, M. Q. (2001). *Qualitative evaluation and research methods* (3rd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Richards, J. C. (2002). Longman Language Teaching and Applied Linguistics. Pearson Education.

Rubadeau, Z. (2016). *An exploration of English language teacher educators' cognitions and practices in relation to the pedagogical purposes and efficacies of 21st-century digital technologies.* Durham theses, Durham University. Available at Durham E-Theses Online: http://etheses.dur.ac.uk/11506/

Saldana, J. (2011). *Fundamentals of qualitative research*. OUP USA.

Seawright, J. (2016). *Multi-methodsocialscience: Combiningqualitativeandquantitativetools*. Cambridge University Press.

Sessoms, D. (2008). Interactive instruction: Creating interactive learning environments through tomorrow's teachers. *International Journal of Technology in Teaching and Learning, 4*(2), 86-96.

Schmidt, D.A., Baran, E., Thompson, A.D., Mishra, P., Koehler, M.J., & Shin, T.S. (2009). Technological pedagogical content knowledge (TPACK): The development and validation of an assessment instrument for pre-service teachers. *Journal of Research Technology in Education, 42*(2), 123–149.

Shin, T., Koehler, M., Mishra, P., Schmidt, D., Baran, E., & Thompson, A. (2009, March). Changing technological pedagogical content knowledge (TPACK) through course experiences. *In Society for Information Technology & Teacher Education International Conference*, 4152-4159. Association for the Advancement of Computing in Education (AACE).

Tai, S. J. D. (2015). From TPACK-in-action workshops to classrooms: CALL competency developed and integrated. *Language Learning & Technology, 19*(1), 139-164.

Teijlingen, van, E., & Hundley, V. (2001). The importance of pilot studies. *Social Research Update*, 35. Retrieved May 2, 2020, from http://www.soc.surrey.ac.uk/sru/SRU35.html.

Tokmak, H. S., Incikabi, L., & Ozgelen, S. (2013). An investigation of change in mathematics, science, and literacy education pre-service teachers' TPACK. *The Asia-Pacific Education Researcher, 22*(4), 407-415.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and teacher education, 17*(7), 783-805.

Tseng, J. J., Cheng, Y. S., & Yeh, H. N. (2019). How pre-service English teachers enact TPACK in the context of web-conferencing teaching: A design thinking approach. *Computers & Education, 128*, 171-182.

Turgut, Y. (2017a). Tracing pre-service English language teachers' perceived TPACK in sophomore, junior, and senior levels. *Cogent Education, 4*(1368612), 1–20.

Turgut, Y. (2017b). A comparison of pre-service, in-service and formation program for teachers' perceptions of technological pedagogical content knowledge (TPACK) in English language teaching (ELT). *Educational Research and Reviews, 12*(22), 1091-1106.

Vereshchahina, T., Liashchenko, O., & Babiy, S. (2018). English language teachers' perceptions of hybrid learning at university level. *Advanced Education, 5*(10), 88-97.

Young, J. R., Young, J. L., & Hamilton, C. (2013). The use of confidence intervals as a meta-analytic lens to summarise the effects of teacher education technology courses on preservice teacher TPACK. *Journal of Research on Technology in Education, 46*(2), 149-172.

Wang, A. Y. (2016, June). TPACK assessment in English language arts for teachers of English as a foreign language. *In EdMedia+ Innovate Learning*,1082-1087. Association for the Advancement of Computing in Education (AACE).

Weerakanto, P. (2019). *Digital literacies of English language teachers and students and their perceptions of technology-enhanced language learning and teaching in Thailand* (unpublished doctoral dissertation). The University of Arizona. Retrieved from https://repository.arizona.edu/handle/10150/633068.

Wetzel, K., & Marshall, S. (2011). TPACK goes to sixth grade: Lessons from a middle school teacher in a high-technology-access classroom. *Journal of Digital Learning in Teacher Education, 28*(2), 73-81.

Winter, G. (2000). A comparative discussion of the notion of validity in qualitative and quantitative research. *The qualitative report, 4*(3), 1-14.

Wu, Y. T., & Wang, A. Y. (2015). Technological, pedagogical, and content knowledge in teaching English as a foreign language: Representation of primary teachers of English in Taiwan. *The Asia-Pacific Education Researcher, 24*(3), 525-533.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 361–378
https://doi.org/10.21449/ijate.780247

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

Factors Affecting Level of Children Resilience and Teachers' Opinions about Resilience

Sibel Yoleri [1,*]

[1] İzmir Democracy University, Department of Preschool Education, İzmir, Turkey

ARTICLE HISTORY

Received: Nov. 23, 2019

Revised: Apr. 2, 2020

Accepted: Aug. 12, 2020

KEYWORDS

Resilience,
Temperament,
Teachers,
Preschool children

Abstract: This research mainly focuses on two purposes, the first of which is to examine the relationship between the resilience levels of 5-6-year-old preschool children, their temperament, and their ages. The second purpose of the research is to determine the opinions of their teachers on resilience and resilient children, the risk factors that affect the resilience and the protective factors. Accordingly, the mixed- method design was used in the study. The sample in the quantitative part of the study consisted of the parents and teachers of the 151 children enrolled in preschool education under the Usak Provincial Directorate for National Education. Qualitative data were collected from the interviews with 15 preschool teachers. The quantitative data were collected using the "Early Childhood Resilience Scale" and "The Short Temperament Scale for Children". The qualitative data were collected using the "Semi-structured Interview Form" which consists of 4 questions regarding the 15 preschool teachers' opinions on resilience. According to the results, the age and temperament (i.e., persistence and reactivity) were found to be significant predictors of resilience. It was also found that the resilience scores of the children increased with age. The qualitative data were analyzed using descriptive and content analysis methods. The teachers expressed the highest rate of resilience as "being able to struggle", while the characteristics of the children, who have resilience behaviour, were described as "being determined". They expressed the concept of "domestic violence" as a risk factor that may influence resilience, and "personality traits" as the protective factor.

1. INTRODUCTION

Research indicates that during early childhood, it is important for children to have a good quality of care and opportunities of learning, adequate nutrition, and community support for families, and also to facilitate the positive development of cognitive, social and self-regulation skills. During these years, the roots of competence are established and many of the most important protective systems for human development emerge. These early years hold great promise for interventions to prevent and reduce risk, boost resources, promote competence and build a strong foundation for future development (Masten, Gewirtz, & Sapienza, 2013). Individuals face with many different situations, changes, positive or negative life events in the

CONTACT: Sibel Yoleri ✉ sibel.yoleri@idu.edu.tr İzmir Democracy University, Education Faculty, Department of Preschool Education, İzmir, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

developmental process and they experience an adaptation process. Various skills and strategies need to be taught as early as possible so that children are prepared for potential adversities and they can make the most of future learning opportunities. The early childhood period is an important stage of life for understanding and promoting resilience. In this process, some individual traits function as the facilitating factors. It is difficult to mention a conventional definition of the concept "resilience", which is referred to in different ways in the literature.

Masten, Best and Garmezy (1990) defined resilience as "individual's having a successful adaptation capacity despite challenging conditions or threats to the development and adaptation of the individual, making efforts to overcome them and ultimately succeeding at". Resilience refers to the ability to overcome challenging situations as well as the ability to be strengthened as a result. Resilience is a developmental and dynamic process (Grotberg, 1995), which is expressed as the ability to recover after the difficulties encountered in individuals' life (Goldstein, & Brooks, 2005), to come up with good results although they encounter risky situations, to get rid of the negative effects of these situations successfully, and more importantly, to be able to revert back to their previous condition (Luthar, 1991; Luthar, Cicchetti, & Becker, 2000; Luthar, 2006; Masten, 2001). Individuals with a high level of resilience are, therefore, able to adapt easily to changing conditions, overcome problems more quickly, and produce solutions to problems in greater numbers and variety (Taylor et al., 2013). At this point, the important issue is seen as being aware of the factors supporting the development of resilience.

Resilience, which is described as the ability of children to overcome social, emotional, developmental, economic, and environmental challenges (Goldstein, & Brooks, 2005), changes depending on innate factors (eg. personality traits such as easy temperament, patience, etc.) and environmental factors (eg. family, school and social environment characteristics) (Masten, & Powell, 2003). Thus, affective, environmental and social characteristics of an individual influence each other and have a common effect on resilience (Hjemdal, 2007; Ungar, 2011, 2012). In many studies, researchers emphasize the ecological approach. Bronfenbrenner's (1994) theory of ecological systems suggests that a person develops in interconnected environments and multiple ecological levels, affecting the development of the individual both directly and through interactions between ecological levels. The ecological system approach refers to the interactions of internal and external forces affecting the behavior of individuals (Danış, 2006; Masten, 2015; Ungar, 2013). This approach draws attention to a variety of factors that shape children's early experiences and influence their levels of resilience (Bronfenbrenner, 1979). According to this perspective, the capacity of the individual to be resilient arises as a result of the level of interactions between personality and environmental factors (Ungar et al., 2007). For example, children who have the advantage of living in a safe community and loving the home environment have greater access to factors that will enable them to exhibit a high level of resistance in the face of adversity (Bowes, Grace, & Hodge, 2012).

Definitions linked to resilience and researches have emphasized two concepts: risk factors and protective factors. It is important that both risk factors and protective factors are referred to at an individual and environmental level. Risk factors are factors that trigger, or cause stress which individuals may encounter. The risk factors, particularly for children, include socio-economic variables (low socio-economic background, poverty, etc.), family variables (negative parental attitudes, separation from parents or having a single parent, death of parents, sick parents, etc.), genetic conditions, child abuse/neglect and negative life experiences (terrorism, immigration, war, natural disasters, etc.) (Greene, 2002; Luthar, Cicchetti, & Becker, 2000; Masten, 2001; Reed-Victor, & Stronge, 2002).

Approaches and skills against risk factors that reduce the effects of the environmental risk or difficulty experienced by the children or allow them to overcome those and improve healthy adaptation are called "protective factors" (Gizir, 2004; Masten, 1994; Sattler, & Font, 2018). Werner and Smith (1992) indicate that protective factors have a significant impact on child development (cit.: Ersay & Erdem, 2017). Protective factors may be found at individual, family and community levels (Wright, Masten, & Narayan, 2013). Protective factors thought to offset the debilitating effects of multiple stress factors in childhood were divided into three categories by Garmezy (1985). This trio of factors has been supported by subsequent studies as well. These factors include; (1) positive temperament, marked self-esteem, ability, and social responsiveness; (2) a supportive family environment that includes a solid relationship with at least one parent; and (3) social support in a non-family environment, such as school or community. In the literature, for example, positive personality traits are listed in the category of individual protective factors (see Smith, & Prior, 1995). In addition, some factors such as intelligence, problem-solving skills, temperament, self-regulation skills based on temperament, coping skills, and social competence are also defined as protective factors (Afifi, & MacMillan, 2011; Benzies, & Mychasiuk, 2009; Lee, & Stewart, 2013; Masten, Best, & Garmezy, 1990; Masten, 2001; Oades-Sese, & Esquivel, 2006). It is important that protective factors outweigh the impact of risk factors that may be exist in children's close surroundings because protective factors can moderate the effects of different risks (Sattler, & Font, 2018).

Temperament is referred to as one of the individual traits that could increase resilience (Compas et al., 2001; Rutter, 1987). Various definitions of temperament among protective factors have been made in studies about resilience. The temperament is the individual differences (Sanson, & Rothbart, 1995), which are biologically based, representing the differences in individual's relativity and self-control (Rothbart, & Bates, 2006), relatively persistent (Sanson, Hemphill, & Smart, 2004), but may vary depending on the stimuli and expectations' change from the environment. Prior et al. (2011) described the temperament as a 'behavior'. Various temperament traits have been expressed in order to reveal the behavior of individuals. Approach/withdrawal, persistence/patience, adaptability, rhythmicity, activity level, intensity of responses, stimulation threshold, distractibility and attention span are some of them (Akın Sarı, 2018; Grist, & McCord, 2010; Yağmurlu, & Kodalak, 2010). Individuals are divided into three groups according to their behaviors they have exhibited since birth, with easy temperament, difficult temperament and slow to warm up temperament. Easy tempered, which is also included among the protective factors, refers to calm, warm-hearted, and cheerful children who can easily adapt to changes. Difficult tempered babies are easy to cry, hard to calm and cannot easily adapt to change. On the other hand, individuals who are slow to warm up tempered are those reacting less negatively compared to difficult tempered children, but sometimes more aggressive (Afifi, & MacMillan, 2011; Thomas, & Chess, 1977; Yağmurlu, & Kodalak, 2010).

It is clear that early childhood is an important time frame for understanding and encouraging, empowering resilience. These early years are important for attempts to prevent and reduce risk, increase resources, increase competence, and build a strong foundation for future development. It is, therefore, necessary to identify risk factors and protective factors in children's lives in order to understand how to develop resilience and to support children. Teachers, as important adults in children's lives, play a significant role in supporting resilience (Hart et al., 2004; Hattie, & Gan, 2011). Children, as role models, when learning about personal feelings, make decisions, share their thoughts, and can help solve problems. (Nolan, Taket, & Stagnetti, 2014). Considering previous studies, although the crucial role of teachers in promoting resilience is highly stated, there are limited studies about preschool teachers' opinions about resilience (Brooks, 2006; Gilligan, 2000; Miljevic-Riđički, Bouillet, & Cefai, 2013).

It is thought that it is important to examine protective factors that support the resilience in the preschool period in order that children can adapt to the challenging and stressful situations they face in their lives. The aim of this study is to determine the predictability of preschool children's age and their temperament traits on resilience. Secondly, it is to evaluate the teachers' opinions and knowledge about being resilient which is an important element in the development and support of resilience. In order to achieve these objectives, the research seeks to answers the following questions:

Is there a relationship between preschool children’s resilience and temperament traits and their ages?

Do the age and temperament traits of preschool children measure children's resilience?

What is the knowledge level of preschool children's teachers about resilience, what are their opinions on characteristics of resilient children, risk factors that can negatively affect children's lives and how to protect them from these factors?

2. METHOD

This section includes research design, sampling study group, data collection tools, data collection and data analysis.

2.1. Research design of the study

In this study, mixed-method research design that combines both quantitative and qualitative patterns was utilized. Mixed method research allows the researcher to combine both qualitative and quantitative methods, approaches and concepts in a study or consecutive studies and thus to better understand and explain the problems (Büyüköztürk et al., 2011; Creswell, 2013). If researchers want to use a mixed method, they should first determine what the purpose of the research is and then decide the order to collect the quantitative and qualitative data (Creswell, 2013). They will then determine the methodology to offer more space, integrate the data collected by the two approaches and eventually establish a theoretical point of view that will shape the basis of the study (Creswell, 2013; Yıldırım & Şimşek, 2013).

Several mixed-methods have been developed in terms of research designs: consecutive descriptive, consecutive discovery, sequential converter, concurrent triangulation, concurrent nested and concurrent converter (Creswell, Plano Clark, Gutmann, & Hanson, 2003; Hanson et al., 2005; Morse, 2003). This study has used a concurrent nested design, in which both qualitative and quantitative data are collected and analyzed simultaneously. Although the quantitative and qualitative data are collected at the same time in the concurrent nested pattern, either quantitative data or qualitative data take up a larger part of the study (Creswell, 2013). Data analysis was conducted separately, and data were combined during interpretation. The data obtained by the quantitative method (Short Temperament Scale for Children, Early Childhood Resilience Scale) were higher and were supported by qualitative method (voice/video recording, semi-structured interview form).

2.2. Participants

As specified by Kemper et al. (2003), the Sequential Quantitative-Qualitative technique is the most commonly used one in the literature. In many studies conducted with this technique, the final sample used in the quantitative stage is employed as a determinant for sampling in the later qualitative stage.

In mixed-method research, sampling refers to sampling and environment selection processes and methods for each of the quantitative and qualitative research (Creswell, 2013). Due to the mixed method, the study group selection was carried out in two stages as both quantitative and qualitative data collection methods were used together.

For the quantitative part of the study, the parents and children's teachers who work in kindergartens in Uşak city center are the target population of the study. The sample consists of the parents and teachers of the children of 5 kindergartens in Uşak city center, who are thought to represent the reachable population and randomly selected from reachable population. During the formation of the sample study group, it was taken into consideration that all children showed a normal development. 49.7% (*n = 75*) of the children were female and 50.3% (*n = 76*) were male. 35.8% (*n = 54*) of the children, who participated in the study, were 5 years old and 64.2% (*n = 97*) were 6 years old.

Fifteen teachers took part in the qualitative aspect of the study. Participants were the preschool teachers who were selected by means of purposive sampling method within the scope of the sample, where quantitative data were obtained. Most qualitative researches do not make a limitation by giving certain numbers; however, 20 to 30 participants in the theory-building studies; in a case study, 4 to 10 participants can be used (Creswell, 2013). 15 of the participating teachers were 20-40 years old. Twelve of these teachers were four-year faculty graduates. Three teachers had an associate degree. Nine of the teachers had 10-15 years of professional experience, four of them had 5-10 years and two of them had 2-3 years. The average number of children in their class was 24.

2.3. Instruments

Under this heading, the tools employed to collect quantitative and qualitative data and the purpose of use of these tools are given.

2.3.1. *Quantitative Data Collection Tools Used in Research*

In the study, the *Demographic Information Form*, *Early Childhood Resilience Scale* and S*hort Temperament Scale for Children* were employed to collect quantitative data.

2.3.1.1. *Demographic information form*

In the form where questions related to personal details were included, the age and the gender of the child, the age and the gender of the parent, their educational status, economic status and the number of children were asked. In the form prepared for teachers, questions about the gender, age, education level, professional experience of the teacher and the number of children in their class were included.

2.3.1.2. *Early childhood resilience scale*

The first form of the scale developed by E. Ersay (individual interview, March 28, 2018) consisted of 51 items. In later analysis, items with a factor load value of less than 0.45 and a factor load value of less than 0.10 (12 items) were excluded from the scale. After the analysis repeated in this direction, it was determined that 39 items showed a single factor structure. The alpha coefficient of the Cronbach's answers given to the 39 items in the final form of the scale was calculated as 0.977. This scale was filled by the teachers for each child selected in compliance with the purpose of the research. To determine the reliability of responses to scale items, the alpha coefficient of Cronbach was estimated from the internal coefficients of consistency, and the alpha coefficient of the Cronbach was defined as 0.942.

2.3.1.3. *The Short temperament scale for children*

This scale (Prior, Sanson, & Oberklaid, 1989) was developed to determine children's temperament characteristics. The scale consists of 30 items with four subscales. Sample items for dimensions were: Reactivity (e.g. 'When upset or annoyed with a task, my child throws it down, cries slams doors, etc.'), Persistence (e.g. "My child is unwilling to leave a game or activity that he/she has not completed"), Rhythmicity (e.g. "My child would like to grab a bite to eat almost at the same time everyday"), Approach/withdrawal (e.g. "My child is shy when first meeting new children"). The internal consistency scores for the original version of the scale were 0.66 for approach, 0.75 for inflexibility/reactivity, 0.75 for persistence, and 0.51 for rhythmicity (Prior, Sanson, & Oberklaid, 1989). In Yağmurlu and Sanson's study (2009) internal consistency was .80 for Approach/Withdrawal, .77 for Reactivity, .48 for Rhythmicity and .76 for Persistence. In this current study, the Cronbach's alpha coefficient scores for Approach subscale was .64, .68 for Reactivity, .65 for Persistence, and .54 for Rhythmicity.

2.3.2. *Qualitative Data Collection Tool Used in Research*

In the qualitative aspect of the study, a case study of qualitative research methods was employed. The interviews is one of the most frequently used data collection tools in qualitative research. Various interview techniques are used in qualitative research (Yıldırım, & Şimşek, 2013). In this study, semi-structured interview technique was used because of the aim of determining the opinions and practices of teachers about resilience and for the flexibility provided by the method.

2.3.2.1. *Semi-Structured interview form*

A semi-structured interview form was prepared to complete the mixed method research, and preschool teachers were asked to share their views on what resilience is, what resilient children's characteristics are, what the risk and protective factors can be. In the preparation of the interview questions developed by the researchers, attention was given to the principles in that to be easy to understand and not to be multidimensional, and that it should not direct the interviewer (Yıldırım & Şimşek, 2013). In the preparation of the form used in the study, the opinions of two faculty members, who are experts in the field having research on qualitative studies, and the information in the related literature were employed. The prepared draft form was submitted to the opinion of two different faculty members, who had studies on preschool education, before performing the trial practice. In order to test the comprehensibility and conformity of the questions with the purpose, the preliminary practice was carried out with two teachers outside the study group. As a result of these interviews, it was determined that there was no problem in terms of comprehensibility and started to work with the working group.

2.4. Procedure

Under this heading, the collection process of quantitative and qualitative data is included.

2.4.1. *Quantitative Data Collection and Research Process*

The temperament traits of the children were filled by the mothers of the children in the sample group. The Early Childhood Resilience Scale was filled by the teachers of the same children.

In the process of collecting quantitative data, firstly permission was obtained from Uşak Provincial Directorate of National Education. Then, in line with the permission, teachers and managers were informed about the study and it was decided to reach parents with training schools and to take advantage of parents' meetings held within the scope of family participating activities. At the end of the meetings, the parents were informed about the scope, purpose and measurement tools of the study and it was explained that the data obtained from the study would be used only within the scope of scientific research, in which their personal details would be kept confidential. Data collection tools did not contain any personal information about either a mother or a child. Only parents who volunteered to participate in the study were included in the study. The measurement tool was sent to 250 families in 5 preschools in Uşak city center, and the total number of completed items was 165 at the end of the data collection process. As a result of the analysis of the collected data, 14 data were excluded due to the lack of information and calculations were made with 151 data sets. In order to make a reliable interpretation, return rate of the measurement tool is recommended to be over 70-80% (Büyüköztürk et al., 2011). It is seen that the ratio in the study is sufficient in this sense.

2.4.2. *Collection of Qualitative Data*

Semi-structured interview form was conducted on 15 volunteer participants among the teachers participating in the research. The interviews lasted approximately 20-30 minutes and were conducted in a relatively quiet area of the school. The interview started with the introduction of interviewer before the questions, the subject of the research was reminded, and its purpose was stated. Then some brifed information was given about the principles of confidentiality. The responses of the teachers, who accepted, were recorded with a voice recorder. The answers to those who did not approve were recorded in writing. In the interview, open-ended and easy-to-understand questions were asked in a certain order.

2.5. Data Analysis

The research study continued from September 2018 to November 2018 for the entire process of data collection. Both quantitative and qualitative data are analyzed at the same time. As in a concurrent qualitative–nested quantitative study, the quantitative data are the primary data resource whereas the qualitative data are supportive of the explanations.

2.5.1. *Analysis of Quantitative Data*

The quantitative aspect of the study was carried out based on the screening model. The screening model, which is one of the descriptive research types and which uses questionnaires or scales as data collection tools, enables the researcher to describe the current situation. In the screening model, participants in a sample from a population are presented with a pre-determined set of questions (Büyüköztürk et al., 2011; Karasar, 2012). If the sample number is less than 30, the Shapiro-Wilk normality test is applied; if it is 30 and more, the Kolmogorov-Smirnov normality test is applied (Büyüköztürk, 2015). Kolmogorov-Smirnov normality test was applied since a total of 151 scales were included in the analysis. As a result of the Kolmogorov-Smirnov normality test, a normal distribution of data was observed.

The correlation coefficient was investigated to determine the relationship between the resilience levels of children and their ages and temperament traits (approach/withdrawal, persistence, rhythmicity, reactivity). The correlation coefficient is used to find and interpret the amount of the relationship between the two variables (Büyüköztürk, 2015). Multiple Regression Analysis was employed to determine whether their ages and temperament traits predicted their resilience and if any, to calculate the predictive power.

2.5.2. *Analysis of Qualitative Data*

For the analysis of the qualitative data obtained from the research, the content analysis was applied on the qualitative data. The main purpose of the content analysis is to reach the concepts and relations that can explain the qualiatively collected data. Within the context of content analysis, the stages such as categorization of the data, finding the themes, arranging and defining the data according to codes and themes, and interpreting the findings follow each other (Yıldırım, & Şimşek 2013). Firstly, the interviews were transformed into a written form by the researchers on computer and tables were formed based on the opinions of the participants. The content analysis continued by reviewing the written data. In the data examined, remarkable and important aspects were determined, followingly codes and then categories were obtained. The code and categories were then made clear by comparing the code and categories produced separately. In order to reflect the opinions of teachers, direct quotations were made from the statements of the teachers. The opinions of the participants were transferred on the basis of confidentiality and coded without giving their names. According to this, teachers were coded as "T" and each participant was given a number as "T1-T15" next to their code.

At the end of the research, two child development specialists, two preschool education specialists and a measurement and evaluation specialist examined the conformity of the responses given, to the themes obtained, during the Validity Reliability Determination Phase. In order to determine the reliability of the study, "consensus" and "dissidence" numbers were determined and used to provide the consistency of judgement across various viewers (inter-rater reliability) suggested by Miles and Huberman's (1994). In qualitative studies, a significant reliability is obtained in cases where the calculation is 70% or higher. Since the reliability of the coding is determined as 82%, it is accepted that the study is reliable (Miles, Huberman, & Saldana, 2014).

3. RESULT / FINDINGS

3.1. Analysis of quantitative data

In the study, the resilience of children was accepted as dependent variable and this variable was tested with multiple regression model to determine how this variable predicted the age and temperament traits of the child. In the research, sub-factors of age and temperament were evaluated together, and progressive multiple regression model was preferred. Firstly, the assumptions required to make the multi-connection model were evaluated. Assuming that the tolerance values are not less than .05 with the assumption that all independent variables are not above .70, the hypotheses that the VIF value is below 10 and that there is no autocorrelation, and that the variables are usually distributed, one by one is evaluated, and the hypothesis that there are no multiple correlations assured.

Correlations related to the relationship between resilience levels, age of children and temperament traits, mean and standard deviation values are shown in Table 1.

Table 1. *Arithmetic Mean, Standard Deviation and Correlation Coefficients of Variables (N=151)*

	1	2	3	4	5	6
Resilience	1	.01	.34**	.12	-.26**	.31**
Approach/Withdrawal		1	.01	-.04	-.07	-.00
Persistence			1	.19**	-.24**	.09
Rhythmicity				1	-.19*	.09
Reactivity					1	.01
Child age						1
$\bar{X}$:	162.46	26.53	27.95	28.92	26.57	1.64
Ss	25.86	6.25	6.09	5.41	7.32	.48

** $p < .01$, * $p < .05$

As seen in Table 1, the correlation analysis revealed no correlation between resilience and being approach/withdrawal and rhythmicity, which are among the temperament traits ($p > .05$). On the other hand, it was observed that there was a statistically significant positive correlation between persistence and temperament traits and resilience ($r = .34$, $p < .01$) and statistically significant negative correlation with reactivity ($r = -. 28$, $p < .05$). When Table 1 is examined, it is observed that there is a statistically significant positive correlation between resilience and age of the child ($r = .31$, $p < .01$).

In the second stage of the analysis of quantitative data, progressive multiple regression analysis was applied to determine whether the temperament traits and the children ages predicted the level of children's resilience, if so, to what extent. The results of the progressive multiple regression analysis are given in Table 2.

Table 2. *Results of progressive multivariate regression analyses*

	β	SHb	β	t	F	R	R^2	ΔR^2
(Constant)	121.60	9.33		13.024^{**}	20.057	.344	.119	.113
Persistence	1.46	.32	.344	4.478^{**}				
(Constant)	99.42	10.64		9.342^{**}				
Persistence	1.36	.31	.320	4.339^{**}	18.301	.445	.198	.187
Child age	15.23	3.97	.283	3.834^{**}				
(Constant)	122.84	13.71		8.957^{**}		.484	.234	.219
Persistence	1.15	.31	.272	3.648^{**}				
Child age	15.62	3.89	.291	4.009^{**}	14.997			
Reactivity	-.69	.26	-.196	2.631^{**}				

Dependent Variable: Resilience, $^{**}p<0.01$

In Table 2, when the R^2 values were examined, it was observed that persistence scores, one of the temperament traits, alone accounted for 12% of the variance [F (1, 149): 20.057; *p*<.01]. Then, when the age of the child was added, it was seen that the persistence and child age explained 20% of the total variance [F (1, 148): 18.301; *p*<.01]. In the final stage, the reactivity score, one of the temperament traits, was added, and they were observed to account for 22% of the variance. According to the t-test results for the independent variables, the age of the child (β = .291; *p*<.01) is the strongest predictor of resilience, and it was followed by persistence, one of the temperament traits (β =.272; *p*<.01) and reactivity, another one of the temperament traits, (β = -.196; *p*<.01). In addition, regression equation shows that the reactivity, one of the temperament traits, expressed negatively the resilience levels, on the contrary, it reflected significantly positively the persistence, one of the temperament traits, and the age of the child. The relationship between resilience and temperament traits (the persistence and the reactivity) and child's age was shown in Figure 1.

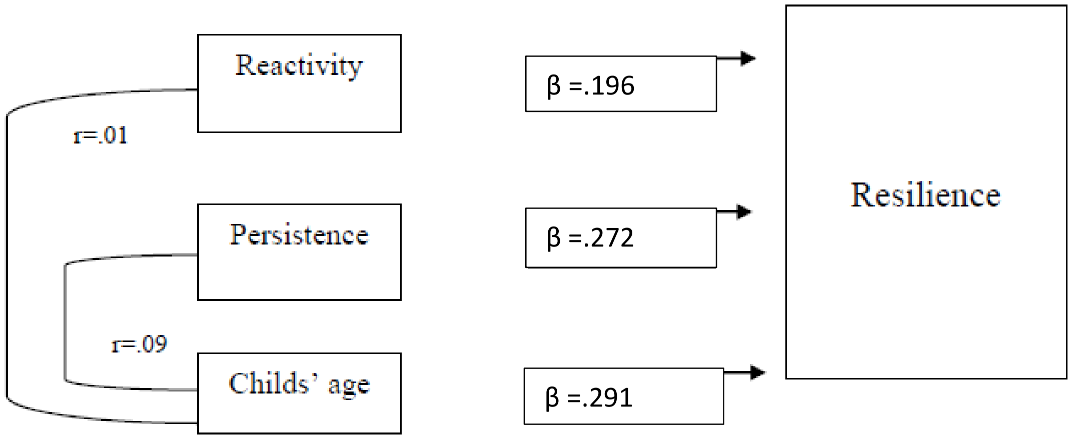


Figure 1. *Model showing the Relation between Resilience and the Variables of "Persistence" and "Reactivity" of Temperament Traits and "the Child's Age"*

3.2. Analysis of qualitative data

In Table 3 the frequency (f) according to the responses given by the preschool teachers, who took part in the research, to the question "What is resilience?". According to the opinions, most ability to struggle (f = 11) was expressed. The teachers' answers to the question included the following:

(T2) "*…despite the difficult conditions not to self-surrender…*"
(T3) "*…resistance to positive or negative situations…*"

(T4) "…*to struggle with the difficulties in achieving the goal*…"
(T11) "…*to struggle, not to give up*…"

Table 3. *Teachers' Opinions on the Concept of "Resilience"*

Themes: What is resilience?		f
	Ability to struggle	11
Codes	Ability to recover one-self	8
	Ability to resist difficulties	5
	Ability to own manage emotions	5
	Not to self-surrender	2
	Determination	2

Table 4. *Teachers' Opinions on Resilient Children's Traits*

Themes: Resilient Children's Traits		f
	To be determined	15
Codes	To be able to direct their attention to different tasks	12
	Ambition	11
	Obstinacy	7
	To have faith in succeeding	7
	To be persistent	7
	To be curious	5
	To be patient, to try till end	5

Table 4 indicates the frequencies (f) of the features expressed by teachers with regard to the features of resilient children. It has been stated that it is the most being determined (f=15) of the opinions. Later, teachers were able to direct the attention of resilient children to different activities (f=12), ambition (f=11), obstinacy (f=7), believing that they could succeed (f=7), be persistent (f=7), be curious (f=5) and stating that they were children who were patient and carried out the activity to the end (f=5). Some of the participants' opinions are as follows;

(T1) "…*to be able to finish a task without getting bored, without giving up*…"
(T3) "…*they do not give up, they try anyway to achieve what they want*…"
(T8) "…*works hard to achieve what he/she wants*…"
(T15) "…*they are confident children*…"

Table 5. *Teachers' Opinions on Risk Factors*

Themes: Risk Factors		f
	Domestic violence	15
Codes	Abuse	15
	Negative financial conditions (eg. poverty)	9
	Parents' attitudes	9
	Death of one of the family members	3
	Technology such as Internet, computer etc.	3

As seen in Table 5, in the study, teachers defined mostly "domestic violence" among the risk factors that may cause resilience. In addition, the risk factors stated by the teachers are abuse (f= 15), negative financial conditions (f= 9), parental attitudes (f= 9), death of one of the family members (f= 3) and the effect of technological devices such as the Internet and computer (f= 3). Some of the participants' opinions are as follows;

(T8) "…*domestic violence is the most important risk factor in my opinion…*"

(T10) "…*children who experienced mother-father death, or their separation are under the risk…*"

(T12) "…*nowadays, I think the computer, internet, tv negatively affect children of all ages…*"

Table 6. *Teachers' Opinions on Protective Factors*

Themes: Protective Factors		f
	Personality traits	13
Codes	Family support	8
	Teachers' approach	5
	School-family cooperation	5

In Table 6, teachers mostly stated "personality traits" among the protective factors in the lives of individuals. Some of the participants' opinions are as follow;

(T4) "…*the support of the family is very important…*"

(T5) "…*not only the family but also the support of other family elders (such as grandparents is very important…*"

(T8) "…*approach of teachers in school…*"

(T14) "…*some child personality traits (eg. temperament, some children very impatient*) …"

4. DISCUSSION and CONCLUSION

The aim of this study was to investigate the relationship between the resilience traits and age and temperament traits of the 5-6-year-old children having preschool education and to investigate the perceptions of preschool teachers' resilience. For this purpose, data were collected from mothers of children having preschool education and from preschool teachers.

In the quantitative aspect of the study, a positive correlation was found between the resilience levels of the children and the children's age and "persistence" among temperament traits, and a negative correlation between "reactivity" among temperament traits. In addition, it was determined that the child's age and persistence and reactivity dimensions of the temperament were predictive variables of child resilience.

Children with persistence temperament traits have the ability to concentrate on a task and organize it. Therefore, these children can develop a positive and optimistic point of view for the future. This ability can help them to cope with negative emotions and positively affect resilience. A significant correlation was found between persistence temperament trait and children's resilience levels (Hutchinson, Stuart, & Pretorius, 2010; Bayındır, Önder, & Balaban Dağal, 2016). However, the results of a study conducted in Turkey show that persistence and reactivity among temperament trait are associated with preschool children's resilience levels (Önder, Balaban Dağal, & Bayındır, 2018). Oades-Sese and Esquivel (2006) studied on the resilience with 207 Afro-American children in 50 economically disadvantaged early childhood classes. Cognitive ability, temperament, autonomy and language skills were found to be protective factors in their studies (Ersay & Erdem, 2017). On the other hand, reactivity refers to being ready to respond to a particular stimulus or event, and this trait being higher makes it difficult to control emotion regulation and behavior in children. Studies show that children with high reactivity experience more externalization problems (Kochanska & Knaack, 2003; Oldehinkel et. al., 2004; Spinrad et al., 2007; Yoleri, 2014). As a finding of the analysis, reactivity as the characteristic of the children's personality decreases, resilience scores decrease, and resilience scores improve if they decrease. This result underpins abstract theories and literature analyses. Reactivity temperament trait was found to be associated with the resilience levels of children (Cumberland-Li, Eisenberg, & Reiser, 2004; Eisenberg et al., 2004; Eisenberg, Spinrad, & Morris, 2002). Similarly, in the literature, individuals with low resilience levels show more problems with inward and outward orientation (Eisenberg, Spinrad, & Morris, 2002; Eisenberg et al., 2010; Kabasakal & Arslan, 2014; Kim & Im, 2014).

As a result of the research, the age of the children was found as a predictor of the resilience of children. When the literature about resilience-age correlation is examined, different results are revealed. Review showed typically that older generation has higher resilience (Campbell-Sills et al., 2009; Herrman et al., 2011; Lundman et al., 2007). It has been emphasized that children at little ages are more and easily vulnerable to all risk factors compared to adolescents and youngsters (Luthar, 1999; as cited in Gizir, 2007). A study by Bayındır et al. (2017) found that 6-year-old children had higher emotion regulation skills than 5-year-olds. According to the teacher evaluation, in a study that examined the resilience levels of preschool children, teachers stated that children's resilience levels of seven-year-old children were higher than the six-year-olds and the five-year-olds were higher than the four-year-olds (Miljević-Riđički, Plantak, & Bouillet, 2017). On the other hand, the findings of this study differ from previous research findings showing that resilience does not change according to age. In the study conducted by Balaban Dağal and Bayındır (2018), a statistically significant result was not found when the resilience level of the children was evaluated in terms of their ages. In another study conducted by Metin (2010), it was indicated that the age did not predict the emotion regulation skill in children of 3-6 age group. In a meta-analysis study, there was no increase in resilience scores as children's ages increased (Nasvytiene, Lazdauskas, & Leonavičiene, 2012).

In the qualitative dimension of the research, in the interviews with the preschool teachers, the questions of "What is resilience?", "What are the characteristics of resilient children?", "What are the risk factors on children and protective factors of resilience?" were asked. In line with these headings, the related themes revealed. Teachers expressed the concepts of ability to struggle (n = 11), self-recovery (n = 8) regarding the concept of resilience. These statements were followed by the ability to resist difficulties (n = 5), ability to own manage emotions (n = 5), not to self-surrender (n = 2), and determination (n = 2). The results of various studies have shown that individuals with high level of resilience are individuals with high levels of self-sufficiency, ability to adapt to changing conditions, ability to change behavior when needed, and problem-solving skills (Taylor et al., 2013). In this sense, the teachers' thoughts on the definition of infidelity are consistent with the literature. In relation to the characteristics of resilient children, teachers stated as to be stable (n = 15), able to direct their attention on different tasks (n = 12), ambition (n = 11), obstinacy, to have faith in succeeding and to be persistent (n = 7), to be curious and behave patiently (n = 5). Their opinions on risk factors included domestic violence (n =15) and abuse (n = 15), negative financial conditions (n = 9), parental attitudes (n = 9), death of one of the family members, and the internet, computer, and so on (n = 3). Personality traits (n = 13) were the first in terms of protective factors in children's lives against risk factors, while it was followed by family support (n = 8), teachers' approach (n = 5) and school-family cooperation (n = 5). Preschool teachers, as the first teacher of children, have a unique opportunity to create a positive effect on the lives of children in preschool classrooms with the idea that every moment of the day is an important moment to increase the resilience of children. Research has also shown that teachers offer positive role models in the lives of flexible children (Cairone, & Mackrain, 2012). Therefore, it is very important to discover the thoughts of teachers about what this phenomenon of resilience means. Ogelman (2015) reveals the relationship between the level of love and warmth of mothers and fathers and the children's resilience. As the level of love and warmth of the parents' increases, the children's resilience increases. Then, the absence of violence in the family, positive behavior in the family, the family environment, the positive perspective of the family events, harmony within the family, raising awareness of the family, and the characteristics of educated parents are in the opinions. In a study by Oswald, Johnson, and Howard (2003), teachers were asked about the factors affecting the development of resilience in students. As a result of the research, the teachers stated that the students' personal inclinations and character traits were the most effective factor in the development of resilience. Green, Oswald, and Spears (2007) asked 14 teachers how they defined resilience and what practices they carried out to support the development of resilience in children. At the end of the study, it was determined that most of the teachers had no accurate information about the resilience and the characteristics of the resilient children. In addition, it was seen that teachers did not consider the concept of risk when explaining the resilience. In the study conducted by Miller-Lewis et al., (2013), they collected information from families and teachers of 485 children between 3-5-years old. It was tried to determine the internal and external forces of the children which can be seen as the protector against the risks. Internal strengths include self-sufficiency, self-esteem, and self-control, while external strengths include relationships between parents and teachers, socio-economic status, family relationships, and stressful life events. In a qualitative study by Miljevic-Riđički, Bouillet, & Cefai (2013), preschool teachers working in Croatian preschools and families were asked questions about resilience and the factors they thought were important for improving children's resilience. The teachers defined the characteristics of the resilient child as self-confident, emotionally mature children. Resnick and Taliaferro (2011) stated that strengthening protective factors could be provided by teachers. Sun and Stewart (2007) stated that school support is important, while Benard (2004) states that teachers have the potential to increase resilience in children through a classroom environment in which children's safety and love and belonging needs are met.

Understanding how children and adolescents growing with the pressure of stressful life experiences will endure of remaining given all the adverse consequences that affect their survival can shed more light on prevention measures for other children and adolescents at comparable risk. Studies have shown that resilience is a personality trait to learn and develop (Bonanno, 2004; Masten, 2001). In this sense, early intervention programs can improve the resilience of children in preschools.

In future studies, longitudinal studies can be suggested to determine the factors affecting the resilience of individuals of different age groups. Moreover, a study on cultural protective factors can be planned.

It would be useful to increase the knowledge about resilience of teachers, who have an important role among the external support systems that increase resilience and help people to overcome the difficulties and to inform them about how they can help when they encounter children who have experienced different risks in their classes by giving training about risk factors.

There are some limitations in this study. Information on resilience was obtained only from teachers. A research on family expectations of resilience is scheduled for the next phase of the study. Another limitation of the study is that the data collected reflect only a cross-section of the time when the data collected. Data on the age and temperament traits of children and their resilience levels can be discussed in detail in longitudinal studies.

Declaration of Conflicting Interests and Ethics

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

ORCID

Sibel Yoleri https://orcid.org/0000-0002-7802-2352

5. REFERENCES

Afifi, T.O., & MacMillan, H. (2011). Resilience following child maltreatment: A review of protective factors. *Canadian Journal of Psychiatry, 56*(5), 266-272.

Akın Sarı, B. (2018). Temperament features and it's impacts on development. *Child Psychiatry-Special Topics, 4*(1), 5-9.

Bayındır, D., Önder, A., & Balaban Dağal, A. (2016). Temperament and resiliency as predictor factors of preschoolers' school readiness. *X. European Conference on Social and Behavioral Science*, Sarajevo, Bosnia-Herzegovina, 19-22 May 2016.

Bayındır, D., Güven, G., Sezer, T., Akşin-Yavuz, E., & Yılmaz, E. (2017). The relationship between maternal acceptance-rejection levels and preschoolers' social competence and emotion regulation skills. *Journal of Education and Learning, 6*(2), 305-316.

Benard, B. (2004). *Resiliency: What we have learned*. San Francisco, CA: WestEd Regional Educational Laboratory.

Benzies, K., & Mychasiuk, R. (2009). Fostering family resiliency: A review of the key protective factors. *Child & Family Social Work, 14*(1), 103-114.

Bonanno, G.A. (2004). Loss, trauma, and human resilience: Have we underestimated the human capacity to thrive after extremely aversive events? *American Psychologist, 59*(1), 20-28.

Bowes, J., Grace, R., & Hodge, K. (2012). *Children, families and communities: Contexts and consequences.* South Melbourne: Oxford University Press.

Brooks, J.E. (2006). Strengthening resilience in children and youths: Maxi-mizing opportunities through the schools. *Children and Schools, 28*, 69-76.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by design and nature.* Cambridge, MA: Harvard University Press.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö.E., Karadeniz, Ş., & Demirel, F. (2011). *Bilimsel araştırma yöntemleri*. Ankara: Pegem Yayınları.

Büyüköztürk, Ş. (2015). *Sosyal bilimler için veri analizi.* Ankara: Pegem Yayınları.

Cairone, K.B., & Mackrain, M. (2012). *Promoting resilience in preschoolers: A strategy guide for early childhood professionals* (2nd ed.). Villanova, PA: Devereux Foundation.

Campbell-Sills, L., Forde, D.R., & Stein, M.B. (2009). Demographic and childhood environmental predictors of resilience in a community sample. *Journal of Psychiatric Research, 43*(12), 1007–1012.

Compas, B.E., Connor–Smith, J.K., Saltzman, H., Thomsen, A.H., & Wadsworth, M.E. (2001). Coping with stress during childhood and adolescence: problems, progress, and potential in theory and research. *Psychological Bulletin, 127*(1), 87–127.

Creswell, J.W., Plano Clark, V. L., Gutmann, M., & Hanson, W. (2003). Advanced mixed methods research designs. In A. Tashakkori, & C. Teddlie (Ed.), *Handbook of mixed methods in social and behavioural research* (pp. 209-240). Thousand Oaks, CA: Sage.

Creswell, J.W. (2013). *Research design qualitative, quantitative, and mixed methods approaches* (4th ed.). London: Sage Publications Inc.

Cumberland-Li, A., Eisenberg, N., & Reiser, M. (2004). Relations of young children's agreeableness and resiliency to effortful control and impulsivity. *Social Development*, *13*(2), 193–212.

Dağal, A.B., & Bayındır, D. (2018). The investigation of the level of ego resilience of preschool children. *Journal of Early Childhood Studies*, *2*(1), 132-150.

Danış, M.Z. (2006). Davranış bilimlerinde ekolojik sistem yaklaşımı. *Aile ve Toplum, 3*(9), 45-53.

Eisenberg, N., Spinrad, T.L., & Morris, A.S. (2002). Regulation, resiliency, and quality of social functioning. *Self and Identity*, *1*(2), 121–128.

Eisenberg, N., Spinrad, T.L., Fabes, R.A., Reiser, M., Cumberland, A., Shepard, SA., Valiente, C., et al. (2004). The relations of effortful control and impulsivity to children's resiliency and adjustment. *Child Development*, *75*(1), 25-46.

Eisenberg, N., Haugen, Rg., Spinrad, T.L., Hofer, C., Chassin, L., & Zhou, Q., et al. (2010). Relations of temperament to maladjustment and ego resiliency in at-risk children. *Social Development*, *19*(3), 577-600.

Ersay, E., & Erdem, E. (2017). *Okul öncesi eğitime devam eden 4-5 yaşındaki çocukların yılmazlık özellikleri ve yılmazlığı destekleyici faktörlerin incelenmesi* (Unpublished masters thesis). Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Gilligan, R. (2000). Adversity, resilience and young people: The protective value of positive school and spare time experiences. *Children and Society, 14*, 37-47.

Gizir, C.A. (2004). *Academic resilience: An Investigation of protective factors contributing to the academic achievement of eighth grade students in poverty* (Unpublished Doctoral Thesis), Middle East Technical University, Ankara.

Gizir, C.A. (2007). A literature review of studies on resilience, risk, and protective factors. *Turkish Psychological Counseling and Guidance Journal, III* (28),113-128.

Goldsmith, H.H., Buss, A.H., Plomin, R., Rothbart, M.K., Thomas, A., & Chess, S. (1987). What is temperament? Four approaches. *Child Development, 58*, 505- 529.

Goldstein, S., & Brooks, R.B. (2005). Why study resilience. In S. Goldstein, & R.B. Brooks (Eds), *Handbook of resilience in children* (pp. 3-15). NY: Springer.

Greene, R.R. (Ed.). (2002). *Resiliency: An integrated approach to practice, policy, and research.* Washington, DC: National Association of Social Workers Press.

Green, D., Oswald, M., & Spears, B. (2007). Teachers' (mis) understandings of resilience. *International Education Journal, 8*(2), 133-144.

Grist, C.L., & McCord, D.M. (2010). Individual differences in preschool children: Temperament or personality. *Infant and Child Development, 19*, 264–274.

Grotberg, E.H. (1995). *A guide to promoting resilience in children: Strengthening the human spirit* (Ed). Bernard van Leer Foundation. Retrieved from https://bibalex.org/baifa/Attachment/Documents/115519.pdf

Hanson, W.B., Creswell, J.W., Plano Clark, V.L., Petska, K.S., & Creswell, D. (2005). Mixed methods research designs in counseling psychology. *Journal of Counseling Psychology, 52*(2), 224-35.

Hart, S., Dixon, A., Drummond, M.J., & McIntyre, D. (2004). *Learning without limits*. Maidenhead: Open University Press.

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In R. Mayer, & P. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 249-271). New York, NY: Routledge.

Herrman, H., Stewart, D.E., Diaz-Granados, N., Berger, E.L., Jackson, B., & Yuen, T. (2011). What is Resilience? *The Canadian Journal of Psychiatry, 56(*5), 258-265.

Hjemdal, O. (2007). Measuring protective factors: The development of two resilience scales in Norway. *Child and Adolescent Psychiatric Clinics of North America, 16*(2),303-321.

Hutchinson, A.K., Stuart, A.D., & Pretorius, H.G. (2010). Biological contributions to well-being: The relationships amongst temperament, character strengths and resilience. *Journal of Industrial Psychology*, *36*(2), 1-10.

Kabasakal, Z., & Arslan, G. (2014). The relationship between psychological resilience, family problems and antisocial behaviors in adolescence. *International Journal of Family, Child and Education, 2*(3), 76-90.

Karasar, N. (2012). *Bilimsel araştırma yöntemleri*. Ankara: Nobel Yayın Dağıtım.

Kemper, E., Stringfield. S., & Teddlie, C. (2003). Mixed methods sampling strategies in social science research. In A. Tashakkori, & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 273-296). Thousand Oaks, CA: Sage.

Kim, D.H., & Im, Y.J. (2014). Resilience as a protective factor for the behavioral problems in school-aged children with atopic dermatitis. *Journal of Child Health Care*, *18*(1), 47-56.

Kochanska, G., & Knaack, A. (2003). Effortful control as a personality characteristic of young children: Antecedents, correlates, and consequences. *Journal of Personality*, *71*(6), 1087-1112.

Lee, P.C., & Stewart, D.E. (2013). Does a socio-ecological school model promote resilience in primary schools? *Journal of School Health, 83*(11), 795-804.

Lundman, B., Strandberg, G., Eisemann, M., Gustafson, Y., & Brulin, C. (2007). Psychometric properties of the Swedish version of the Resilience Scale. *Scandinavian Journal of Caring Sciences, 21*(2), 229-37.

Luthar, S.S. (1991). Vulnerability and resilience: A study of high-risk adolescents. *Child Development, 62*(3), 600-616.

Luthar, S.S., Cicchetti, D., & Becker, B. (2000). The construct of resilience: A critical evaluation and guidelines for future work. *Child Development, 71*(3), 543–562.

Luthar S.S. (2006). Resilience in development: A synthesis of research across five decades. In: D. Cicchetti, & D.J., Cohen (Eds.), *Developmental psychopathology (2nd ed., vol.3: Risk, disorder, and adaptation* (pp.739-795). New York: Wiley.

Masten, A.S., Best, K.M., & Garmezy, N. (1990). Resilience and development: Contributions from the study of children who overcome adversity. *Development and Psychopathology, 2*(4)*,* 425-444.

Masten, A.S. (1994). Resilience in individual development: Successful adaptation despite risk and adversity: Challenges and prospects. In M. Wang, & E. Gordon (Eds.), *Educational resilience in inner city America: Challenges and prospects* (pp. 3-25). Hillsdale, NJ: Lawrence Erlbaum.

Masten, A.S. (2001). Ordinary magic: Resilience processes in development. *American Psychologist*, *56*(3), 227-238.

Masten, A.S., & Powell, J.L. (2003). A resilience framework for research, policy, and practice. In S.S. Luthar (Ed.) *Resilience and vulnerabilities: Adaptation in the context of childhood adversities.* New York: Cambridge University Press.

Masten, A.S., Gewirtz, A.H., & Sapienza, J.K. (2013). Resilience in development: The importance of early childhood. In R.E. Tremblay, R. G. Barr., & R. DeV. Peters (Eds.), *Encyclopedia on early childhood development* [online]. Retrieved from http://www.child-encyclopedia.com/sites/default/files/textes-experts/en/834/resilience-in-development-the-importance-of-early-childhood.pdf

Masten, A.S. (2015). Pathways to integrated resilience science. *Psychological Inquiry*, *26*(2), 187–196.

Metin, İ. (2010). *The effects of dispositional anger, effortful control and maternal responsiveness on turkish preschoolers' emotion regulation* (Unpublished master's thesis). Koç Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Miljević-Riđički, R., Bouillet, D., & Cefai, C. (2013). Pre-curriculum activities: Focus groups on resilience. In Bertram, T., Formosinho, J., Lohmander, M.K., Veisson, M., Ugaste, A., Õun, T., Tuuling, L. (Eds.), *Abstract Book of the Value, Culture and Contexts 23rd Annual ECERA Conference* (p.52), Tallinn University, Tallinn. Retrieved from: https://www.eecera.org/wp-content/uploads/2013/08/2013-tallinn.pdf

Miljević-Riđički, R., Plantak, K., & Bouillet, D. (2017). Resilience in preschool children–the perspectives of teachers, parents and children. *International Journal of Emotional Education*, *9*(2), 31-43.

Miller-Lewis, L.R., Searle, A. K., Sawyer, M.G., Baghurst, P.A., & Hedley, D. (2013). Resource factors for mental health resilience in early childhood: An analysis with Multiple methodologies. *Child and Adolescent Psychiatry and Mental Health, 7*(6), 1-23.

Miles, M.B., Huberman, A.M., & Saldana, J. (2014). *Qualitative data analysis: A methods sourcebook and the coding manual for qualitative researchers* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Morse, J.M. (2003). Principles of mixed methods and multimethod research design. In A. Tashakkori, & C. Teddlie. (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 189-208). Thousand Oaks, CA: Sage Publications.

Nasvytiene, D., Lazdauskas, T., & Leonavičiene, T. (2012). Child's resilience in face of maltreatment: A meta-analysis of empirical studies. *Psichologija, 46*, 7-26.

Nolan, A., Taket, A., & Stagnitti, K. (2014). Supporting resilience in early years classrooms: The role of the teacher. *Teachers and Teaching: Theory and Practice, 20*(5), 595-608.

Oades-Sese, G.V., & Esquivel, G.B. (2006). Resilience among at-risk Hispanic American preschool children. *Annals of the New York Academy of Sciences*, *1094*(1), 335-339.

Ogelman, H.G. (2015). Predictor effect of parental acceptance-rejection levels on resilience of preschool children. *Social and Behavioral Sciences, 174,* 622-628.

Oldehinkel, A.J., Hartman, C., DeWinter, A.F., Veenstra, R., & Ormel, J. (2004). Temperament profiles associated with internalizing and externalizing problems in preadolescence. *Development and Psychopathology*, *16*(2), 421–440.

Oswald, M., Johnson, B., & Howard, S. (2003). Quantifying and evaluating resilience-promoting factors: Teachers' beliefs and perceived roles. *Research in Education, 70*(1), 50-64.

Önder, A., Dağal, A. B., & Bayındır, D. (2018). The predictive effect of preschool children's temperament characteristics and parenting styles of mothers on ego resiliency level of children. *Education & Science*, *43*(193), 79-90.

Prior, M., Sanson, A., & Oberklaid, F. (1989). The Australian Temperament Project. In G.A. Kohnstammve J.E. Bates, & M.K. Rothbart (Eds.), *Temperament in childhood* (pp. 537-556). Chichester: John Wiley and Sons.

Prior, M., Bavin, E., Cini, E., Eadie, P., & Reilly, S. (2011). Relationships between Language Impairment, Temperament, Behavioural Adjustment and Maternal Factors in A Community Sample of Preschool Children. *International Journal of Language & Communication Disorders, 46*(4), 489-494.

Reed-Victor, E., & Stronge, J. H. (2002). Homeless students and resilience: Staff perspectives on individual and environmental factors. *Journal of Children Poverty*, *8*(2), 159-183.

Resnick, M.D., & Taliaferro, L.A. (2011). Resilience. In B. Bradford Brown & M. Prinstein (Eds.), *Encyclopedia of adolescence* (pp. 299-306). San Diego, CA: Academic Press.

Rothbart, M.K., & Bates, J.E. (2006). Temperament. In W. Damon, R.M. Lerner., & N. Eisenberg (Eds.), *Handbook of child psychology, Six edition*: *Social, emotional, and personality development* (pp. 99-166). New York: John Wiley & Sons Inc.

Rothbart, M.K. (2011). *Becoming who we are: Temperament and personality in development*. New York, NY: The Guilford Press.

Rutter, M. (1987). Psychosocial resilience and protective mechanisms. *American Journal of Orthopsychiatry, 57*, 316–331.

Sanson, A., & Rothbart, M.K. (1995). Child temperament and parenting. In M. Bornstein (Ed.), *Handbook of Parenting* (Vol:4, pp. 299-321). Hillsade, NJ: Erlbaum.

Sanson, A., Hemphill, S.A., & Smart, D. (2004). Connections between temperament and social development: A review. *Social Development, 13*, 142–170.

Sattler, K.M.P., & Font, S.A. (2018). Resilience in young children involved with child protective services. *Child Abuse & Neglect*, *75*, 104-114.

Smith, J., & Prior, M. (1995). Temperament and stress resilience in school-age children: A within-families study. *Journal of the American Academy of Child & Adolescent Psychiatry, 34*(2), 168–179.

Spinrad, T.L., Eisenberg, N., Gaertner, B., Popp, T., Smith, C.L., Kupfer, A., Greving, K., Liew, J., & Hofer, C. (2007). Relations of maternal socialization and toddlers' effortful control to children's adjustment and social competence. *Developmental Psychology*, *43*(5), 1170–1186.

Sun, J., & Stewart, D. (2007). Age and gender effects on resilience in children and adolescents. *International Journal of Mental Health Promotion, 9*(4), 16-25.

Taylor, Z.E., Eisenberg, N., Spinrad, T.L., & Widaman, K.F. (2013). Longitudinal relations of intrusive parenting and effortful control to ego-resiliency during early childhood, *Child Development*, *84*(4), 1145-1151.

Thomas, A., & Chess, S. (1977). *Temperament and development*. New York: Brunner/Mazel.

Ungar, M., Brown, M., Liebenberg, L., Othman, R., Kwong, W.M., Armstrong, M., & Gilgun, J. (2007). Unique pathways to resilience across cultures. *Adolescence, 42*(166), 287–310.

Ungar, M. (2011). The social ecology of resilience: Addressing contextual and cultural ambiguity of a nascent construct. *American Journal of Orthopsychiatry, 81*(1), 1-17.

Ungar, M. (Ed.) (2012). *The social ecology of resilience: A handbook of theory and practice.* New York, NY: Springer.

Ungar, M. (2013). Resilience after maltreatment: The importance of social services as facilitators of positive adaptation. *Child Abuse & Neglect, 37*(2-3), 110–115.

Wright, M.O.D., Masten, A.S., & Narayan, A.J. (2013). Resilience processes in development: Four waves of research on positive adaptation in the context of adversity. In S. Goldstein & R. B. Brooks, *Handbook of resilience in children* (2nd ed., pp. 15–37). New York, NY: Springer.

Yıldırım, A., & Şimşek, H. (2013). *Sosyal bilimlerde nitel araştırma yöntemleri*. Ankara: Seçkin Yayınları.

Yağmurlu. N., Sanson. A., & Köymen. B. (2005). Effects of parenting and child temperament on the development of prosocial behavior: The mediating role of theory of mind. *Turkish Journal of Psychology, 20*(55), 1-20.

Yağmurlu, B., & Sanson, A. (2009). Parenting and temperament as predictors of prosocial behavior in Australian and Turkish Australian children. *Australian Journal of Psychology*, *61*, 77-88.

Yağmurlu, B., & Kodalak, A.C. (2010). Bağlanma, mizaç ve ebeveyn-çocuk ilişkileri [Attachment, temperament and parent-child relationships]. (T. Solmuş (Ed.), Bağlanma, Evlilik ve Aile Psikolojisi içinde [In attachment, marriage and family psychology], (111-125). İstanbul: Sistem.

Yoleri, S. (2014). The relationship between temperament, gender, and behavioural problems in preschool children. *South African Journal of Education*, *34*(2), 1-18.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 379–391
https://doi.org/10.21449/ijate.684360
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Selection of Scholarship Students in Higher Education with VIKOR Method

Kubra Akilli [1], **Emre Ipekci Cetin** [2,*]

[1]Marmara University, School of Banking and Insurance, Department of Actuary, Istanbul, Turkey
[2]Akdeniz University, Faculty of Economics and Administrative Sciences, Department of Econometrics, Antalya, Turkey

ARTICLE HISTORY

Received: Feb. 02, 2020
Revised: July 3, 2020
Accepted: Aug 14, 2020

KEYWORDS

Scholarship Student Selection,
Higher Education,
Multi Criteria Decision Making,
VIKOR Method

Abstract: Selection of students who will benefit from scholarships given in the university are usually done by formed commission. Due to limited number of scholarships offered, commission are obliged to choose the most appropriate students. In this selection process, it is important to make objective evaluation. The commission should mostly interview the applicants face to face. This situation causes time and labour loss and a stressful environment for both members of the commission and the students. An objective scoring system could solve the problems discussed above. In this study, 200 students who applied for the scholarship at Akdeniz University Faculty of Economics and Administrative Sciences to the scholarship were ranked. In this study, firstly the selection criteria of students for the scholarship was determined with the help of researchers and social aid service experts. Then, the weights of the criteria were calculated by the SWING method. These weights were used to rank the students who were eligible for the scholarship by using the VIKOR method. This method will make an objective evaluation and will accelerate the selection process.

1. INTRODUCTION

Defined as unrequited assistance to successful and needy students, the scholarship supports students in meeting their physiological and cultural expenses such as accommodation, nutrition, transportation and education. Institutions and organizations select students for scholarship by using various evaluation criteria. Applications are generally evaluated by the commission which formed by these institutions and organizations and the students to be awarded scholarships are determined. Limited number of scholarships makes hard the selection of appropriate student for the commission. Selecting students to be awarded a scholarship from candidate students is a complex decision-making process that requires multiple selection criteria to be considered simultaneously. In this respect, it would be appropriate to approach the scholarship selection process as a multi-criteria decision-making problem.

Many problems may have more than one qualitative or quantitative, contradictory criterion and purpose. One alternative may be best for one criterion, while it may be worse for another criterion. Multi-criteria decision making (MCDM) is a part of operations research that supports

CONTACT: Emre İPEKÇİ ÇETİN ✉ ecetin@akdeniz.edu.tr Akdeniz University, Faculty of Economics and Administrative Sciences, Department of Econometrics, Antalya, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

the decision maker to resolve problems when multiple conflicting criteria are involved and need to be evaluated (Sitorus, Cilliers, & Brito-Parada, 2019). It assists the decision-maker in finding a best choice to these situations.

Multi-criteria decision-making problems are grouped under three headings: Selection, Sorting, and Classification problems. In selection problems, the aim is to determine the best alternative. In the ranking problems, it is aimed that the alternatives will be defined correctly or measurably from good to bad. In classification problems, alternatives are classified according to a preference or criterion. (Yıldırım & Önder, 2015). This study is a ranking problem applied on to scholarship student selection.

There are various studies using MCDM methods on to student selection problems. For example, Yeh (2003) formulated the scholarship student selection as Multiattribute decision making and used comparative methods including Total Sum Method, Simple Additive Weighting (SAW), the Weighted Product (WP) and TOPSIS. Altunok, Özpeynirci, Kazançoğlu and Yılmaz (2010) discussed three MCDM methods namely Analytic Hierarchy Process (AHP), Weighted Product (WP) and TOPSIS method for postgraduate student selection. Mavrotas and Rozakis (2012) proposed PROMETHEE V2 method for selection of students for a postgraduate program. Taşkın, Üstün, and Deliktaş (2013) ranked candidate students for Erasmus Student Mobility by Fuzzy AHP method. Mahmud, Pazil, Mazlan, Jamaluddin, and Hasan (2017) applied Fuzzy AHP to selection of eligible students in receiving the scholarship while Irvanizam (2018) applied Fuzzy TOPSIS method. Deliktaş and Üstün (2017) handled the student selection process in the Erasmus program. They proposed an integrated approach of fuzzy Multimoora and Multichoice Conic Goal Programming. De Farias Aires, Ferreira, Araujo, and Borenstein (2017) developed a hybrid algorithm called ELECTRE-TOPSIS for rank students in Brazilian University. Mardhiyyah, Sejati, and Ratnasari (2019) used MOORA method as decision support system selection process for scholarship selection.

Besides the above studies there are various studies about scholarship selection by using MCDM in Turkey. For example, Erdem Hacıköylü (2006) used AHP to determine the students who will receive nutrition and shelter assistance from Anadolu University. Criteria are grouped into the income status of the family, student's success, student accommodation and the number of children, the presence of parents and siblings' education. By the AHP method, the students who were eligible for help were compared. Abalı, Kutlu, and Tamer (2012) handled the problem of selecting a student for a scholarship at Kırıkkale University Faculty of Engineering. The criteria are the number of children depend on the family, the total monthly income of the family, the status of the parents, the total number of properties owned by the family and the employment status of the student. As a result of the AHP, it was determined that the most important criterion was the total monthly income of the family. By TOPSIS method the most appropriate student for the scholarship was chosen among the five students. Çakır (2016) handled the problem of determining the students at Adnan Menderes University Nazilli Faculty of Economics for part-time job by using AHP based VIKOR method. The main criteria for ranking the students are academic qualification, the monthly income of the student, the number of dependents of the family, the status of the parents, the total monthly income of the family and the family assets. The weights of the criteria were determined by AHP and the student's monthly income was found as the most critical criterion. With the VIKOR method, the 448 applicants were ranked, and first 50 students were invited for interview. Pençe, Tarhan, and Çetinkaya Bozkurt (2017) handled student selection problem for Turkey Education Foundation scholarship at Mehmet Akif Ersoy University Faculty of Education. The criteria are age, gender, class, number of courses failed, OSYM ranking, parental status, the number of dependents of the family, the annual income of the family and the status of the property of the family. As a result of AHP, the criteria with the highest weight was annual income of the student's family. At the end of the study, 27 applicants were ranked by using the TOPSIS method and the first three candidate were found eligible for the scholarships.

The most important part of the scholarship selection process is to objective evaluation of the candidates. An objective scoring system could provide decision support to the commission for selecting appropriate students for scholarships. For this purpose, in this study MCDM based scoring system is proposed for an objective and compromised selection process.

2. METHOD

This study was conducted at the beginning of the 2017-2018 academic year, using the information given by 200 students who were studying at Akdeniz University Faculty of Economics and Administrative Sciences. In this study, firstly, the criteria affecting the selection of students for scholarship were determined. The importance weights of criterion calculated by using SWING method. Then, the candidate students were ranked by using the VIKOR method which is one of the multi-criteria decision-making method. Weights of criteria were used in VIKOR method as an input. Since the simplicity and the flexibility of use and understandable procedure makes the VIKOR method suitable for this ranking problem regarding the scholarship students. The VIKOR method was preferred in this study because it is an effective tool for multi-criteria decision making, especially in a situation where the decision maker cannot express or know its preference at the beginning of the system design. This method offers compromise solutions for problems related to conflicting criteria, focusing on raking and selecting a range of specific alternatives.

2.1. VIKOR Method

VIKOR method focuses on ranking and selecting from a set of alternatives and determines a compromise solution for a problem with conflicting criteria, which can help the decision-makers to reach a final decision. Here, the compromise solution is a feasible solution, which is the closest to the ideal, and a compromise means an agreement established by mutual concessions. The method provides a maximum group utility for the majority and a minimum of an individual regret for the opponent. It determines the compromise ranking list and compromises the solution by introducing the multi-criteria ranking index based on the particular measure of closeness to the ideal solution. This ranking index is an aggregation of all criteria, the relative importance of the criteria, and a balance between total and individual satisfaction (Liu, Mao, Zhang & Li. 2013). VIKOR method has been applied in many different fields such as supplier selection (Alimardani, Zolfani, Aghdaie, & Tamosaitiene, 2013; Fei, Deng, & Hu, 2019; Abdel-Baset, Chang, Gamal, & Smarandache, 2019), performance evaluation (Kumar, Aswin, & Gupta, 2020; Ture, Dogan, & Kocak, 2019; Buyukozkan & Karabulut, 2017; Wu, Lin, & Chang, 2011; Rezaie, Ramiyani, Nazari-Shirkouhi, & Badizadeh, 2014; Ranjan, Chatterjee, & Chakraborty, 2016; Kaya, İpekçi Çetin, & Kuruüzüm, 2011; Chen & Chen, 2010), personnel selection (Krishankumar, Premaladha, Ravichandran, Sekar, Manikandan, & Gao, 2020), service quality (Gupta, 2018; Yang, Su, & Wang, 2017; Lin, Chen, Chuang, & Lin, 2016), material selection (Jahan, Mustapha, Ismail, Sapuan, & Bahraminasab, 2011; Dev, Aherwar, & Patnaik, 2020) .

Assuming that the rows in the decision matrix represent the alternatives and the columns represent the criteria, the solution steps of the VIKOR method continue as follows (Opricovic & Tzeng, 2004; Büyüközkan & Ruan, 2008; Tong, Chen, & Wang, 2007; İpekçi Çetin & Çetin, 2016; Paksoy, 2017; Çetin & İpekçi Çetin, 2010):

Step 1. Determination the best f_i^* and the worst f_i^- values of all criterion functions, i=1, 2,…,n. If the i-th function represents a benefit, then

$$f_i^* = \underset{j}{max} f_{ij} \quad f_i^- = \min_j f_{ij} \quad \text{if the i-th function represents a benefit;}$$

$$f_i^* = \min_j f_{ij} \quad f_i^- = \max_j f_{ij} \quad \text{if the i-th function represents a cost.} \tag{1}$$

Step 2. Computation the values S_j and R_j, j=1, 2,…, J

$$S_j = \sum_{i=1}^{n} w_i (f_i^* - f_{ij}) / (f_i^* - f_i^-) , \tag{2}$$

$$R_j = \max_i [w_i (f_i^* - f_{ij}) / (f_i^* - f_i^-)] , \tag{3}$$

Here w_i are the weights of criteria.

Step 3. Computation the values Q_j, j=1, 2… J

$$Q_j = v(S_j - S^*) / (S^- - S^*) + (1-v)(R_j - R^*) / (R^- - R^*) \tag{4}$$

Where $S^* = \min_j S_j$, $S^- = \max_j S_j$, $R^* = \min_j R_j$, $R^- = \max_j R_j$

v is introduced as weight of the strategy of "the majority of criteria" (or "the maximum group utility"), here $v = 0.5$.

Step 4. Ranking the alternatives, sorting by the values S_j, R_j and Q_j. The results are three ranking lists.

Step 5. Proposing as a compromise solution the alternative (a') which is ranked the best by the measure Q (minimum) if the following two conditions are satisfied:

C1: "Acceptable advantage": $Q(a'') - Q(a') \geq DQ$ Where a'' is the alternative $DQ = 1/(J-1)$; J is the number of alternatives.

C2. "Acceptable Stability in decision making": The alternative a' must also be the best ranked by S or/and R. This compromise solution is stable within a decision-making process, which could be the strategy of maximum group utility (when v > 0.5 is needed), or "by consensus" v ≈ 0.5, or "with veto" (v < 0.5). Here, v is the weight of decision-making strategy of maximum group utility.

If one of the conditions is not satisfied, then a set of compromise solutions is proposed, which consists of:

- Alternatives a′ and a″ if only condition C2 is not satisfied, or
- Alternatives a′,a″,…,a(M) if condition C1 is not satisfied; and a(M) is determined

by the relation Q(a(M))−Q(a′)<DQ for maximum M (the positions of these alternatives are "in closeness").

The best alternative, ranked by *Q*, is the one with the minimum value of *Q*. The main ranking result is the compromise ranking list of alternatives, and the compromise solution with the "advantage rate".

2.1.1. *Weights calculation for criteria*

Weights express the relative importance of criteria. As decision makers expressing the importance of criteria can be supported with several methods such as SWING method, SMART, AHP, MACBETH, PAPRIKA (Pazsto, Jurgens, Tominc, & Burian, 2020; Nemeth, Molnar, Bozoki, Wijaya, Inota, Campbell, & Kalo, 2019). Due to its ease in application and the simplicity of its calculations, the SWING method was selected for determining the weights of the criteria. This method makes it easier and more reliable for researchers to get expert ideas.

In the SWING method, performance measurements are considered to be between 0-100. A score of 100 is given to the most important criterion, and then progress is made by providing a score of less than 100 to other criteria. The decision-maker scores all the criteria according to their importance. Finally, normalization is performed by dividing each score to the sum of all scores (Wang, Jing, Zhang, & Zhao, 2009).

In this study, for determining the criteria weights, scoring was done by six academicians who participated in Akdeniz University Scholarship and Social Services Committee. The weights of each criteria calculated with geometric mean of six scores given by academicians. Final weights presented in Table 1.

Table 1. *Criteria and weight of scholarship selection*

	Criteria effective in selection of scholarship	Weights
C1	Having a martyr relative	0.138
C2	The existence of an individual with disability in the student's family	0.135
C3	Monthly income of student's family	0.125
C4	Monthly income of student	0.110
C5	The number of people the head of the family is responsible for caring	0.096
C6	Type of student's social assurance	0.084
C7	Where the student earns his income	0.082
C8	Whether the place where the family lives is rent	0.059
C9	Student's place of residence	0.043
C10	The residence of the student's family	0.032
C11	Whether parents are alive and their marital status	0.028
C12	The father's profession	0.025
C13	The mother's profession	0.022
C14	Education level of the mother	0.010
C15	Education level of the father	0.010

As it can be seen from Table 1, the criterion of having a martyr relative has the highest weight. Education level of the mother and father are the criteria with the lowest weight criteria that is effective in selecting students to be awarded scholarships.

2.1.2. *Establishment of decision matrix*

The decision matrix consists of 15 criteria and 200 alternatives (students). Students are studying Akdeniz University Faculty of Economics and Administrative Sciences. The values of students for the criteria are obtained from the Scholarship Application Form and Scoring System which created by Social Services. Sample values of data can be seen in Table 2.

Table 2. *Decision matrix*

Weights	0,138	0,135	0,125	0,110	0,096	0,084	0,082	0,059	0,043	0,032	0,028	0,025	0,022	0,010	0,010
Student Number	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
1	10	10	100	80	4	60	40	50	100	40	30	60	0	50	50
2	10	10	100	100	2	40	100	50	40	40	30	30	100	30	20
3	10	10	100	100	6	100	100	0	70	40	0	30	100	40	20
4	10	10	100	100	4	60	60	0	40	40	0	30	100	50	40
5	10	10	100	80	4	60	80	50	40	40	30	60	100	40	40
6	10	10	100	100	5	60	60	0	100	30	0	30	0	40	40
7	10	10	100	60	3	60	60	0	100	40	0	30	100	40	40
8	10	10	100	80	5	60	40	0	80	40	0	100	60	30	30
9	10	10	100	20	0	0	40	0	100	30	0	80	100	40	40
10	10	10	0	80	9	100	40	50	40	40	0	60	0	50	50
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
191	10	10	80	100	5	40	40	50	100	40	0	40	0	40	20
192	10	10	100	80	12	100	100	0	100	40	30	0	0	50	50
193	10	10	100	100	9	100	60	50	80	30	0	80	100	20	20
194	10	10	100	80	4	60	40	0	80	30	0	40	100	40	40
195	10	10	100	60	4	100	60	0	80	40	0	100	100	40	40
196	10	10	100	60	4	60	60	0	80	40	0	30	100	30	30
197	10	10	100	80	3	100	80	50	100	40	30	0	100	40	40
198	10	10	100	60	3	60	80	50	40	40	60	60	0	20	20
199	10	10	100	20	5	60	40	0	100	30	60	0	100	40	40
200	10	10	80	80	6	0	40	50	70	40	0	40	100	20	10

2.1.3. *Calculations of VIKOR Method*

Firstly, the best f_i^* and the worst f_i^- values of all criterion functions are determinate from equation (1). After that with using the equation (2), (3) and (4); Sj, Rj and Qj are calculated for each student j=1,2,…,200. (Qj values are computed by selecting v=0.5). Table 3 and Table 4 gives the S and R scores of students respectively while Table 5 gives Q scores and their corresponding rankings.

The students whose numbers are 119, 44 and 89 have the highest score respectively according to VIKOR method. The student with the lowest score is the student 66.

The best alternative (student) according to the Q-values is the student 119 with the minimum value of Q. It satisfies condition C1 and C2. Because $Q(a'') - Q(a') = 0.180 - 0.166 \geq DQ = 0.005$ and this student is also the best ranked by R. Therefore, student 119 has an acceptable advantage and acceptable stability with respect to the other students.

Table 3. *S scores of students*

Rank	Student No	Si	Rank	Student No	Si	Rank	Student No	Si	Rank	Student No	Si	Rank	Student No	Si
1	89	0.247	41	133	0.446	81	166	0.510	121	8	0.547	161	99	0.580
2	183	0.271	42	3	0.446	82	198	0.511	122	88	0.549	162	20	0.581
3	121	0.316	43	16	0.450	83	184	0.512	123	126	0.549	163	153	0.585
4	163	0.340	44	106	0.452	84	11	0.512	124	39	0.549	164	152	0.586
5	25	0.352	45	162	0.452	85	164	0.513	125	136	0.552	165	170	0.586
6	142	0.353	46	165	0.454	86	195	0.514	126	135	0.552	166	95	0.587
7	129	0.356	47	123	0.455	87	112	0.514	127	81	0.553	167	63	0.588
8	80	0.359	48	23	0.457	88	154	0.514	128	120	0.553	168	77	0.591
9	51	0.366	49	176	0.458	89	177	0.514	129	179	0.554	169	141	0.593
10	104	0.368	50	33	0.458	90	174	0.515	130	19	0.555	170	61	0.596
11	160	0.373	51	134	0.462	91	105	0.516	131	75	0.555	171	199	0.597
12	56	0.377	52	5	0.463	92	90	0.516	132	137	0.555	172	79	0.598
13	193	0.392	53	125	0.463	93	22	0.518	133	30	0.556	173	117	0.607
14	57	0.397	54	53	0.470	94	72	0.522	134	32	0.556	174	158	0.611
15	98	0.401	55	46	0.470	95	127	0.523	135	28	0.559	175	132	0.612
16	60	0.402	56	2	0.471	96	114	0.524	136	35	0.559	176	155	0.615
17	71	0.402	57	27	0.474	97	187	0.525	137	128	0.560	177	140	0.617
18	192	0.415	58	67	0.477	98	24	0.527	138	7	0.564	178	156	0.618
19	50	0.415	59	143	0.479	99	191	0.527	139	74	0.564	179	116	0.619
20	14	0.416	60	21	0.479	100	69	0.527	140	194	0.565	180	65	0.622
21	43	0.416	61	107	0.480	101	161	0.528	141	62	0.565	181	93	0.622
22	18	0.417	62	108	0.482	102	13	0.531	142	103	0.565	182	86	0.623
23	181	0.421	63	94	0.482	103	186	0.531	143	113	0.565	183	169	0.625
24	159	0.422	64	68	0.484	104	47	0.533	144	87	0.567	184	157	0.629
25	119	0.422	65	17	0.484	105	150	0.533	145	64	0.569	185	31	0.629
26	146	0.423	66	109	0.486	106	6	0.535	146	168	0.569	186	178	0.631
27	197	0.426	67	1	0.488	107	78	0.535	147	190	0.569	187	124	0.634
28	173	0.427	68	97	0.488	108	4	0.535	148	196	0.569	188	130	0.640
29	58	0.427	69	48	0.488	109	49	0.537	149	200	0.573	189	110	0.647
30	182	0.433	70	73	0.491	110	38	0.538	150	52	0.573	190	37	0.649
31	44	0.437	71	26	0.497	111	148	0.539	151	172	0.574	191	40	0.652
32	59	0.437	72	144	0.498	112	91	0.539	152	151	0.574	192	180	0.654
33	118	0.438	73	41	0.499	113	189	0.541	153	138	0.575	193	82	0.656
34	70	0.439	74	42	0.503	114	122	0.542	154	29	0.575	194	54	0.658
35	55	0.440	75	145	0.504	115	36	0.543	155	102	0.576	195	34	0.661
36	131	0.441	76	12	0.504	116	147	0.544	156	188	0.576	196	149	0.668
37	15	0.442	77	96	0.505	117	84	0.544	157	85	0.577	197	139	0.674
38	45	0.442	78	92	0.506	118	175	0.544	158	111	0.578	198	9	0.695
39	167	0.443	79	171	0.506	119	115	0.545	159	10	0.579	199	101	0.747
40	185	0.445	80	100	0.509	120	83	0.546	160	76	0.580	200	66	0.774

Table 4. *R scores of students*

Rank	Student No	Ri	Rank	Student No	Ri	Rank	Student No	Ri	Rank	Student No	Ri	Rank	Student No	Ri
1	44	0.135	41	39	0.137	81	80	0.137	121	121	0.137	161	161	0.137
2	119	0.135	42	40	0.137	82	81	0.137	122	122	0.137	162	162	0.137
3	1	0.137	43	41	0.137	83	82	0.137	123	123	0.137	163	163	0.137
4	2	0.137	44	42	0.137	84	83	0.137	124	124	0.137	164	164	0.137
5	3	0.137	45	43	0.137	85	84	0.137	125	125	0.137	165	165	0.137
6	4	0.137	46	45	0.137	86	85	0.137	126	126	0.137	166	166	0.137
7	5	0.137	47	46	0.137	87	86	0.137	127	127	0.137	167	167	0.137
8	6	0.137	48	47	0.137	88	87	0.137	128	128	0.137	168	168	0.137
9	7	0.137	49	48	0.137	89	88	0.137	129	129	0.137	169	169	0.137
10	8	0.137	50	49	0.137	90	89	0.137	130	130	0.137	170	170	0.137
11	9	0.137	51	50	0.137	91	90	0.137	131	131	0.137	171	171	0.137
12	10	0.137	52	51	0.137	92	91	0.137	132	132	0.137	172	172	0.137
13	11	0.137	53	52	0.137	93	92	0.137	133	133	0.137	173	173	0.137
14	12	0.137	54	53	0.137	94	93	0.137	134	134	0.137	174	174	0.137
15	13	0.137	55	54	0.137	95	94	0.137	135	135	0.137	175	175	0.137
16	14	0.137	56	55	0.137	96	95	0.137	136	136	0.137	176	176	0.137
17	15	0.137	57	56	0.137	97	96	0.137	137	137	0.137	177	177	0.137
18	16	0.137	58	57	0.137	98	97	0.137	138	138	0.137	178	178	0.137
19	17	0.137	59	58	0.137	99	98	0.137	139	139	0.137	179	179	0.137
20	18	0.137	60	59	0.137	100	99	0.137	140	140	0.137	180	180	0.137
21	19	0.137	61	60	0.137	101	100	0.137	141	141	0.137	181	181	0.137
22	20	0.137	62	61	0.137	102	101	0.137	142	142	0.137	182	182	0.137
23	21	0.137	63	62	0.137	103	102	0.137	143	143	0.137	183	183	0.137
24	22	0.137	64	63	0.137	104	103	0.137	144	144	0.137	184	184	0.137
25	23	0.137	65	64	0.137	105	104	0.137	145	145	0.137	185	185	0.137
26	24	0.137	66	65	0.137	106	105	0.137	146	146	0.137	186	186	0.137
27	25	0.137	67	66	0.137	107	106	0.137	147	147	0.137	187	187	0.137
28	26	0.137	68	67	0.137	108	107	0.137	148	148	0.137	188	188	0.137
29	27	0.137	69	68	0.137	109	108	0.137	149	149	0.137	189	189	0.137
30	28	0.137	70	69	0.137	110	109	0.137	150	150	0.137	190	190	0.137
31	29	0.137	71	70	0.137	111	110	0.137	151	151	0.137	191	191	0.137
32	30	0.137	72	71	0.137	112	111	0.137	152	152	0.137	192	192	0.137
33	31	0.137	73	72	0.137	113	112	0.137	153	153	0.137	193	193	0.137
34	32	0.137	74	73	0.137	114	113	0.137	154	154	0.137	194	194	0.137
35	33	0.137	75	74	0.137	115	114	0.137	155	155	0.137	195	195	0.137
36	34	0.137	76	75	0.137	116	115	0.137	156	156	0.137	196	196	0.137
37	35	0.137	77	76	0.137	117	116	0.137	157	157	0.137	197	197	0.137
38	36	0.137	78	77	0.137	118	117	0.137	158	158	0.137	198	198	0.137
39	37	0.137	79	78	0.137	119	118	0.137	159	159	0.137	199	199	0.137
40	38	0.137	80	79	0.137	120	120	0.137	160	160	0.137	200	200	0.137

Table 5. *Q scores for v=0.50 and students rankings*

Rank	Student No	Qi	Rank	Student No	Qi	Rank	Student No	Qi	Rank	Student No	Qi	Rank	Student No	Qi
1	**119**	0.166	41	133	0.689	81	166	0.749	121	8	0.785	161	99	0.816
2	**44**	0.180	42	3	0.689	82	198	0.750	122	88	0.786	162	20	0.817
3	**89**	0.500	43	16	0.692	83	184	0.751	123	126	0.787	163	153	0.820
4	183	0.523	44	106	0.694	84	11	0.751	124	39	0.787	164	152	0.821
5	121	0.565	45	162	0.695	85	164	0.753	125	136	0.789	165	170	0.822
6	163	0.588	46	165	0.696	86	195	0.753	126	135	0.790	166	95	0.822
7	25	0.599	47	123	0.697	87	112	0.753	127	81	0.790	167	63	0.823
8	142	0.601	48	23	0.699	88	154	0.753	128	120	0.790	168	77	0.827
9	129	0.604	49	176	0.700	89	177	0.753	129	179	0.791	169	141	0.829
10	80	0.606	50	33	0.700	90	174	0.755	130	19	0.792	170	61	0.831
11	51	0.613	51	134	0.704	91	105	0.755	131	75	0.792	171	199	0.832
12	104	0.614	52	5	0.705	92	90	0.756	132	137	0.792	172	79	0.833
13	160	0.619	53	125	0.705	93	22	0.757	133	30	0.793	173	117	0.842
14	56	0.623	54	53	0.711	94	72	0.761	134	32	0.793	174	158	0.845
15	193	0.638	55	46	0.711	95	127	0.762	135	28	0.796	175	132	0.846
16	57	0.642	56	2	0.713	96	114	0.763	136	35	0.796	176	155	0.849
17	98	0.646	57	27	0.715	97	187	0.763	137	128	0.797	177	140	0.851
18	60	0.647	58	67	0.718	98	24	0.765	138	7	0.801	178	156	0.852
19	71	0.647	59	143	0.720	99	191	0.766	139	74	0.801	179	116	0.853
20	192	0.659	60	21	0.720	100	69	0.766	140	194	0.801	180	65	0.855
21	50	0.660	61	107	0.721	101	161	0.767	141	62	0.802	181	93	0.856
22	14	0.660	62	108	0.722	102	13	0.769	142	103	0.802	182	86	0.857
23	43	0.660	63	94	0.723	103	186	0.769	143	113	0.802	183	169	0.859
24	18	0.661	64	68	0.724	104	47	0.771	144	87	0.803	184	157	0.862
25	181	0.665	65	17	0.725	105	150	0.771	145	64	0.805	185	31	0.863
26	159	0.666	66	109	0.727	106	6	0.773	146	168	0.806	186	178	0.864
27	146	0.667	67	1	0.728	107	78	0.773	147	190	0.806	187	124	0.868
28	197	0.670	68	97	0.729	108	4	0.773	148	196	0.806	188	130	0.873
29	173	0.670	69	48	0.729	109	49	0.775	149	200	0.809	189	110	0.880
30	58	0.671	70	73	0.731	110	38	0.776	150	52	0.810	190	37	0.881
31	182	0.676	71	26	0.737	111	148	0.777	151	172	0.810	191	40	0.884
32	59	0.681	72	144	0.738	112	91	0.777	152	151	0.810	192	180	0.886
33	118	0.681	73	41	0.739	113	189	0.779	153	138	0.811	193	82	0.888
34	70	0.682	74	42	0.743	114	122	0.780	154	29	0.811	194	54	0.890
35	55	0.683	75	145	0.744	115	36	0.781	155	102	0.812	195	34	0.892
36	131	0.684	76	12	0.744	116	147	0.781	156	188	0.812	196	149	0.900
37	15	0.685	77	96	0.745	117	84	0.782	157	85	0.813	197	139	0.906
38	45	0.685	78	92	0.745	118	175	0.782	158	111	0.814	198	9	0.925
39	167	0.686	79	171	0.745	119	115	0.783	159	10	0.815	199	101	0.975
40	185	0.688	80	100	0.749	120	83	0.784	160	76	0.816	200	**66**	**1.000**

3. DISCUSSION and CONCLUSION

In this study, with the help of researchers and social aid service experts, the criteria which must be considered while selecting students for scholarship are determined. Then, these criteria were weighted by the scholarship committee members with SWING method. The criterion of having a martyr relative was found as the most important criterion. The second most important criterion is the presence of a disabled person in the family. The lowest scoring criteria among the 15 criteria are the education level of both the father and mother. Weights which was found by the SWING method were used in the VIKOR method.

According to the results of the VIKOR method, the student in the first place (number 119) stays in a rented house, his/her family lives in the rural area without paying rent. The student has no disability in himself/herself or his/her family but has martyr relative. His/her parents are alive and living together. And his/her father works as a civil servant. The number dependent member of the family is 4.

It was determined that there were only two students who had martyr relationship in their families. The VIKOR method placed these two students in the first two places as this criterion has the highest weight.

It is tried to provide a decision support on student selection for scholarship by using SWING and VIKOR methods in this study. The criteria affecting the selection of student for scholarship were determined with the cooperation of researchers and social aid service experts. If MCDM methods will be used in student selection for scholarship, the determination of criteria and the determination of their weights is the most important part, because results are very sensitive to these parameters. The expertise and number of people whose opinion will be taken in determining the parameters will increase the reliability of the results. So that, by applying more experts in scholarship field may increase the reliability of the study. In this study, the application of integrated VIKOR method recommended to commission to help their decision in student selection for scholarship. Although the proposed system will provide an objective decision mechanism, it cannot be said that it eliminates the need for an interview.

In addition, different multi-criteria decision-making methods can be applied, and the results can be compared. By integrating methods into a computer software, a decision support platform can be developed for the use of commissions.

Declaration of Conflicting Interests and Ethics

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

ORCID

Kübra Akıllı https://orcid.org/0000-0001-5474-3051
Emre İpekçi Çetin https://orcid.org/0000-0002-8108-1919

4. REFERENCES

Abalı, Y. A., Kutlu, BS., & Tamer, E. (2012). Çok ölçütlü karar verme yöntemleri ile bursiyer seçimi [Multicriteria decision making methods with selection of scholarship holder: application in an educational institution]. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi*, *26*(3-4), 259-272.

Abdel-Baset, M., Chang, V., Gamal, A., & Smarandache, F. (2019). An integrated neutrosophic ANP and VIKOR method for achieving sustainable supplier selection: A case study in importing field. *Computers in Industry*, *106*, 94-110.

Alimardani, M., Zolfani, S.H., Aghdaie, M.H., & Tamošaitienė J. (2013). A novel hybrid SWARA and VIKOR methodology for supplier selection in an agile environment, *Technological and Economic Development of Economy*, *19*(3), 533-548

Altunok, T., Özpeynirci, Ö., Kazançoğlu, Y., & Yılmaz, R. (2010). Comparative Analysis of Multicriteria Decision Making Methods for Postgraduate Student Selection. *Eurasian Journal of Educational Research (EJER)*, *10*(40), 1-15.

Büyüközkan, G., & Karabulut, Y. (2017). Energy project performance evaluation with sustainability perspective. *Energy*, *119*, 549-560.

Büyüközkan, G. & Ruan, D. (2008). Evaluation of software development projects using a fuzzy multi-criteria decision approach. *Mathematics and Computers in Simulation*, *77*(5-6), 464-475.

Chen, J. K., & Chen, I. S. (2010). Aviatic innovation system construction using a hybrid fuzzy MCDM model. *Expert Systems with Applications*, *37*(12), 8387-8394.

Çakır, E. (2016). Kısmi zamanlı olarak çalışacak öğrencilerin Analitik Hiyerarşi Prosesi temelli VIKOR yöntemi ile belirlenmesi [The determination of part-time students using Vikor method based on Analytic Hierarchy Process]. *International Journal of Management Economics and Business*, *12*(29), 195-224.

Çetin, M. K., & İpekçi Çetin, E. (2010). Multi-criteria analysis of banks' performances. *International Journal of Economics and Finance Studies, 2*(2), 73-78.

De Farias Aires, R. F., Ferreira, L., de Araujo, A. G., & Borenstein, D. (2017). Student selection in a Brazilian university: using a multi-criteria method. *Journal of the Operational Research Society*, 1-14.

Deliktas, D., & Ustun, O. (2017). Student selection and assignment methodology based on fuzzy MULTIMOORA and multichoice goal programming. *International Transactions in Operational Research*, *24*(5), 1173-1195.

Dev, S., Aherwar, A., & Patnaik, A. (2020). Material selection for automotive piston component using Entropy-VIKOR method. *Silicon*, *12*(1), 155-169.

Erdem Hacıköylü, B. (2006). Analitik Hiyerarşi karar verme süreci ile Anadolu Üniversitesi'nde beslenme ve barınma yardımı alacak öğrencilerin belirlenmesi [Determination of students who will receive nutritional and shelter aid at Anadolu University through analytical hierarchy decision making process]. Master's thesis, Anadolu University, Eskişehir.

Fei, L., Deng, Y., & Hu, Y. (2019). DS-VIKOR: A new multi-criteria decision-making method for supplier selection. *International Journal of Fuzzy Systems*, *21*(1), 157-175.

Irvanizam, I. (2018). Application of the fuzzy topsis multi-attribute decision making method to determine scholarship recipients. *In Journal of Physics: Conference Series,* (978:1), 012056. IOP Publishing.

İpekçi Çetin, E., & Çetin, H. H. (2016). Using VIKOR method for analyzing of qualification levels and transition to employment of European Union and candidate countries. *The Online Journal of Science and Technology, 6*, 99-102.

Jahan, A., Mustapha, F., Ismail, M. Y., Sapuan, S., & Bahraminasab, M. (2011). A comprehensive VIKOR method for material selection. *Materials & Design*, *32*, 1215-1221.

Kaya, P., İpekçi Çetin, E., & Kuruüzüm, A. (2011). Çok kriterli karar verme ile Avrupa Birliği ve aday ülkelerin yaşam kalitesinin analizi [Analysis of quality of life in European Union and candidate countries with multi-criteria decision making]. *İstanbul Üniversitesi İktisat Fakültesi Ekonometri ve İstatistik E- Dergisi*, *13*, 80-94.

Krishankumar, R., Premaladha, J., Ravichandran, K. S., Sekar, K. R., Manikandan, R., & Gao, X. Z. (2020). A novel extension to VIKOR method under intuitionistic fuzzy context for solving personnel selection problem. *Soft Computing*, *24*(2), 1063-1081.

Kumar, A., Aswin, A., & Gupta, H. (2020). Evaluating green performance of the airports using hybrid BWM and VIKOR methodology. *Tourism Management*, *76*, 10394

Lin, C. K., Chen, Y. S., Chuang, H. M., & Lin, C. Y. (2016). Using VIKOR to improve E-service quality performance in E-store. In *Frontier Computing* (pp. 1041-1049). Singapore: Springer.

Liu, HC., Mao, LX., Zhang, ZY., & Li, P. (2013). Induced aggregation operators in the VIKOR method and its application in material selection. *Applied Mathematical Modelling*, *37*(9), 6325-6338.

Mahmud, N., Pazil, N. S. M., Mazlan, U. H., Jamaluddin, S. H., & Hasan, N. N. C. (2018). Scholarship Eligibility and Selection: A Fuzzy Analytic Hierarchy Process Approach. *Proceedings of the Second International Conference on the Future of ASEAN (ICoFA) 2017–Volume 2* (pp. 175-183), Singapore: Springer.

Mardhiyyah, R., Sejati, R. H. P., & Ratnasari, D. (2019). A Decision Support System of Scholarship Grantee Selection Using Moora. *International Journal of Applied Business and Information Systems*, *3*(1), 21-27.

Mavrotas, G., & Rozakis, S. (2009). Application in a Students' Selection Problem. *Journal of Decision Systems*, *18*(2), 203-229.

Németh, B., Molnár, A., Bozóki, S., Wijaya, K., Inotai, A., Campbell, J.D., & Kaló, Z. (2019). Comparison of weighting methods used in multicriteria decision analysis frameworks in healthcare with focus on low-and middle-income countries. *J. Comp. Eff. Res.*, *8*, 195-204.

Opricovic, S., & Tzeng, G. H. (2004). Compromise solution by MCDM methods: a comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research*, *156*(2), 445-455.

Paksoy, S. (2017). *Çok kriterli karar vermede güncel yaklaşımlar [Current approaches in multi-criteria decision making]*. Adana: Karahan Kitabevi.

Pászto, V. Jürgens, C. Tominc, P., & Burian, J. (2020). *Spationomy*. Springer Nature.

Pençe, İ., Tarhan, L., & Çetinkaya Bozkurt, Ö. (2017). Türk Eğitim Vakfı Bursu verilecek uygun adayların AHP ve TOPSIS yöntemi kullanılarak belirlenmesi: Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi örneği [Determination of suitable candidates for Turkish education foundation scholarship by using AHP and TOPSIS Method: Mehmet Akif Ersoy University Faculty of Education example]. *Journal of Applied Sciences of Mehmet Akif Ersoy University*, *1*(1), 37-49.

Ranjan, R., Chatterjee, P., & Chakraborty, S. (2016). Performance evaluation of Indian Railway zones using DEMATEL and VIKOR methods. Benchmarking*: An International Journal*, *23*(1), 78-95.

Rezaie, K., Ramiyani, S. S., Nazari-Shirkouhi, S., & Badizadeh, A. (2014). Evaluating performance of Iranian cement firms using an integrated fuzzy AHP–VIKOR method. *Applied Mathematical Modelling*, *38*(21-22), 5033-5046.

Saat, M. (2000). Çok amaçlı karar vermede bir yaklaşım: Analitik Hiyerarşi Yöntemi [An approach to multi-objective decision making: Analytical Hierarchy Process]. *Gazi University Journal of Economic and Administrative Sciences*, *2*(2), 149-162.

Sitorus, F., Cilliers, J. J., & Brito-Parada, P. R. (2018). Multi-criteria decision making for the choice problem in mining and mineral processing: Applications and trends. *Expert Syst. Appl., 121*, 393–417.

Taşkın, H. Üstün, Ö., & Deliktaş, D. (2013) Fuzzy MCDM approach for oral examination in Erasmus student selection process. *Dumlupınar Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *32*, 21-40.

Tong, L. I., Chen, C. C., & Wang, C. H. (2007). Optimization of multi-response processes using the VIKOR method. *The International Journal of Advanced Manufacturing Technology*, *31*(11-12), 1049-1057.

Türe, H., Doğan, S., & Koçak, D. (2019). Assessing Euro 2020 strategy using multi-criteria decision making methods: VIKOR and TOPSIS. *Social Indicators Research*, *142*(2), 645-665

Wang, J. J., Jing, Y.Y., Zhang, C.F., & Zhao, J.H. (2009). Review on multi-criteria decision analysis aid in sustainable energy decision–making. *Renewable and Sustainable Energy Reviews*, *13*(9), 2263-2278.

Wu, H. Y., Lin, Y. K., & Chang, C. H. (2011). Performance evaluation of extension education centers in universities based on the balanced scorecard. *Evaluation and Program Planning*, *34*(1), 37-50.

Yang, M. H., Su, C. H., & Wang, W. C. (2017). The use of a DANP with VIKOR approach for establishing the model of e-learning service quality. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(8), 5927-5937.

Yeh, C. H. (2003). The selection of multiattribute decision making methods for scholarship student selection. *International Journal of Selection and Assessment*, *11*(4), 289-296.

Yıldırım, B. F., & Önder, E. (2015). Çok kriterli karar verme yöntemleri [Multi criteria decision making methods]. Bursa: Dora Basım.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 392–403

https://doi.org/10.21449/ijate.711073

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Validity and Reliability of The Cancer Loneliness and The Cancer-Related Negative Social Expectations Scale

Ebru Kara [1], **Ilgun Ozen Cinar** [1,*]

[1]Department of Public Health Nursing, Faculty of Health Sciences, Pamukkale University, Denizli, Turkey.

ARTICLE HISTORY

Received: Mar 29, 2020
Revised: May 26, 2020
Accepted: June 20, 2020

KEYWORDS

Cancer,
Loneliness,
Negative Social Expectation,
Validity,
Reliability

Abstract: There is a need for cancer-specific tools to evaluate loneliness and cancer-related negative social expectations before developing interventions for cancer patients. The purpose of this study was to examine the reliability and validity of the Cancer Loneliness and the Cancer-related Negative Social Expectations Scale. Data were collected from 300 cancer patients registered to an oncology outpatient clinic of a University Hospital for this methodological study. In the data collection, Patient Information Form, Cancer Loneliness Scale and Cancer-related Negative Social Expectations Scale and the General Loneliness Scale were used. The Cronbach's Alpha coefficient of the Cancer Loneliness Scale was found to be .88, Spearman-Brown correlation value was found to be .81, CFI, .98, GFI, .96, X^2/SD, 2.99 and RMSEA .08. As for the Negative Social Expectations Scale, Cronbach alpha value was found as .82, Spearman-Brown correlation value .86, CFI 1.00, GFI 1.00, X^2/SD 1.33 and RMSEA .02. The study revealed those both scales were highly reliable and indices of fit showed perfect fit. These scales are highly valid and reliable instruments for the Turkish society.

1. INTRODUCTION

Cancer is a significant reason for morbidity and mortality in all regions and countries. Globally, 18.1 million individuals were diagnosed with cancer and 9.6 million individuals lost their lives due to cancer. Although there are a number of proven interventions to prevent cancer, the promotion and implementation of preventive measures has an important place in this process (Bray et al., 2018).

The aim of cancer diagnosis and treatment programs is to prolong the life of patients and to enable the best possible lives for the survivors (WHO, 2019). Loneliness which is a well-known risk factor for mental and physical health is a negative concept for the health of cancer patients (Jaremka et al., 2013). Cancer patients face some symptoms that are both psychological and somatic. Those people also suffer from anxiety and social difficulties during and after their treatment. Moreover, such an experience makes patients feel lonely (Brintzenhofe-Szoc, Levin, Li, Kissane & Zabora, 2009; Kroenke, Johns, Theobald, Wu & Tu, 2013). In particular, loneliness which decreases immune function, and increases depression in cancer patients may

CONTACT: İlgün ÖZEN ÇINAR ✉ iocinar@pau.edu.tr Department of Public Health Nursing, Faculty of Health Sciences, Pamukkale University, Denizli, Turkey.

increase fatigue, pain, sleep disturbance and cause mortality together with other factors (Jaremka et al., 2014a).

According to the loneliness theory, negative social expectations may cause much more negative relationships, increasing loneliness and the related negative social expectations (Decxk, Akker & Buntinx, 2014). Negative social expectations may specifically be associated with cancer experience. Families and friends of those individuals should provide support and sympathy after the diagnosis. If such behaviours are not seen, then the patients could feel disappointed. Loneliness theory and the relevant studies have shown loneliness can lead to negative expectations in cancer patients (Adams, 2016).

In the care of cancer patients, although loneliness is taken into consideration as a part of care, no effective techniques are found in order to identify and intervene for cancer-related loneliness. In defining cancer-related loneliness, healthcare professionals should trust the patients who express feelings of loneliness or use a variety of approaches to reveal loneliness. However, these approaches are not sufficient to evaluate cancer patients (Wells & Kelly, 2008; Macmillan Cancer Support, 2014).

Studies in the literature, on loneliness in cancer patients generally use UCLA loneliness scale. There have not been many studies related to this issue. Jaremka et al. (2014b) found that for breast cancer survivors, loneliness increases the risk of pain, depression and fatigue symptom cluster and also affects physical and mental health. Fanakidou et al. (2018) found a higher level of loneliness in young breast cancer individuals and in patients without breast reconstruction within one year after mastectomy. In another study it was noted that minimized social support was related to the elevated loneliness and hopelessness (Pehlivan, Ovayolu, Ovayolu, Sevinç, & Camcı, 2012). Dodds et al. (2015) found no difference between loneliness levels in experimental and control groups in breast cancer patients despite educational intervention; whereas in another study it was found that the loneliness level decreased in the experimental group (Tabrizi, Radfar & Taei, 2016). All these studies describe the loneliness of cancer patients in general.

There is a need for cancer-specific tools to evaluate loneliness before developing interventions for cancer patients (Adams, Mosher, Winger, Abonour & Kroenke, 2018). The study was carried out in order to evaluate the reliability and validity of the Cancer Loneliness Scale and the Cancer-related Negative Social Expectations Scale developed by Adams et al.

2. METHOD

2.1. Study Design

The research is a methodological one.

2.2. Setting and Sampling

The population of the study was composed of adult cancer patients admitted to the Oncology Polyclinic of a University for treatment and control purposes. For factor analysis, 200 subjects were considered to be "moderate", 300 subjects were considered "good", 500 subjects were considered "very good", and 1000 subjects were considered to be "excellent" (Streiner & Kottner 2014; Tavşancıl, 2014). In this context, study sample consisted of 300 cancer patients. Inclusion criteria were as follows: Patients diagnosed with cancer in 2016 and 2017 and enrolled in the oncology outpatient clinic, aged 18 years or older, without any communication problems, no brain cancer as primary diagnosis. The participants' consents were obtained as well. Since it was determined that loneliness did not differ according to the type or stage of cancer (Decxk et al., 2014), all types and stages of cancer (except primary brain cancer) were included in the study. Data were obtained from breast, lung, colon, ovarian, prostate, cervical, kidney and pancreatic cancer patients by means of face-to-face interview method between April and August

2018. Six patients who were diagnosed with primary brain cancer were excluded because of impaired perception and comprehension (Adams, 2016).

2.3. Measurements

The following tools were used for data collection.

2.3.1. *Patient Information Form*

It has totally 18 questions including the socio-demographic characteristics of the patients and diagnostic and therapeutic information about their diseases. The questions were formed by the researchers based on the literature (Jaremka et al., 2014a; Tabrizi et al., 2016; Adams et al., 2018).

2.3.2. *Cancer Loneliness Scale* (*CLS*)

Developed by Adams et al. (2017), the original scale included 15 items based on loneliness theory. The scale was later revised into a 7-item one-dimensional form after validity and reliability analyses. The scale is used in cancer patients to evaluate cancer associated loneliness (i.e., attributed loneliness cancer experience). Items are scored as follows: Never (1), rarely (2), sometimes (3), often (4), and always (5). When the score gets higher, it means there will be an increase in terms of cancer associated loneliness. The Cronbach's alpha coefficient of the scale was .94 (Adams et al., 2017; Adams et al., 2018).

2.3.3. *Cancer-Related Negative Social Expectations Scale* (*CRNSES*)

Developed by Adams et al. (2017), the original scale consisted of 14 items based on loneliness theory and previous studies. It was later revised into a 5-item one-dimensional form after validity and reliability analyses. The scale assesses the negative social cognition of the patients about their cancer experiences. The items are scored as follows: strongly disagree (1), partially disagree (2), slightly disagree (3), slightly agree (4), partially agree (5), and strongly agree (6). Cronbach's alpha coefficient was .90 (Adams et al., 2017).

2.3.4. *UCLA Loneliness Scale*

The scale which was developed by Russel, Peplau and Ferguson in 1978 was revised in 1980 by the same authors so that half of the items in the scale were positive and half were negative. The third version of the scale consists of 20 items with 11 negative and 9 positive statements. The reliability of the scale, which exhibits a one-dimensional factor structure, was determined to be between .89 and .94 in studies on different samples (students, nurses, teachers, elderly) (Russell, 1996). UCLA is commonly preferred in order to calculate general loneliness. It is a 4-point Likert type scale with responses between 1 (never) and 4 (always). The reliability and validity study of the Turkish scale was performed by Demir (1988) and alpha reliability coefficient found as .94 (Demir, 1988).

2.4. Cultural Adaptation of Scales

In this study, the recommendations of the World Health Organization (WHO, 2017) and the International Testing Commission (ITC, 2018) reference guidelines, which define the steps to be followed in adaptation studies, were considered in the cultural adaptation of the scales. The guidance published by the International Test Commission is in line with WHO although there may be changes in steps in some cases. The first step is the adaptation of language and culture (WHO, 2017; ITC, 2018). First, permission to use the scales was obtained from the original author via e-mail, and language validity was performed. The scales were translated into Turkish by different health experts whose native language is Turkish and who speak fluent English. The translations were evaluated by the researchers together with a specialist working in the field. The Turkish version of the scales was created through selecting the most appropriate narratives for each item. In the second step, semantic expressions should be considered. The scale, which was evaluated by an expert of Turkish Language and Literature, was finalized after necessary arrangements were made. The third step is the expert panel. In this step, concordance ratio between the opinions of 8 experts was calculated with the Content Validity Index (CVI). Davis method was preffered in CVI calculation. At least 3, at most 20 experts evaluate each item as follows; (a) "highly appropriate" (4 points), (b) "appropriate but minor change" (3 points), (c) "item needs to be revised" (2 points) and (d) "item not suitable" (1 point). In this technique, the number of experts selecting options (a) and (b) is divided by the total number of experts in order to obtain content validity index (CVI). Provided that the CGI index is greater than 0.80, the content validity of the item is considered sufficient (Davis, 1992; Erdoğan, Nahcivan & Esin, 2017). According to the content validity analysis of our study, the intelligibility levels of the items were found to be between .88 and 1.00. In the fourth step, the scales were translated back to English by a professional translator whose native language is English and compared with the original scale by the researchers. Pilot application and cognitive analysis were performed in the fifth step. The scales were administered to 30 cancer patients resembling the sample and all items were understood by the participants. After these steps, the final version of the scales was obtained and the scales were given serial numbers. In the last step, documentation was made and a report was created.

2.5. Ethical Considerations

Permission was received from the Non-Interventional Clinical Ethics Committee of a University (dated 10.01.2018 and numbered 60116787-020/2485). In addition, verbal consent was obtained from the patients together with institutional permission.

2.6. Statistical Analysis

Data were analyzed by SPSS 24 (Statistical Package for the Social Sciences) and Lisrel 8.80 (Linear Structural Relations) statistical software programs. Significance level was taken as $p < 0.05$ in all statistical evaluations. Descriptive statistics were presented as number and percentage. As part of validity analysis, CVI was calculated for content and scope validity. Confirmatory Factor Analysis (CFA) was used for the construct validity of the scales and Pearson Product-Moment Correlation was examined for concurrent validity. In the reliability analysis of the data; normal distribution of the scales was calculated with Skewness and Kurtosis Coefficients, and item analysis, internal consistency and Split-half reliability were evaluated by means of Item Total Score Correlation, Cronbach's Alpha Coefficient and Spearman-Brown Coefficient Value. In this study, since the data is normally distributed, the Maximum Likelihood method was used as the parameter estimation method in CFA.

3. RESULT / FINDINGS

57.0% of the patients were found to be female while 43.0% were found to be male. The mean age was found to be 57.03 ± 11.32 (min.19 - max.84). 27.0% of patients were breast cancer, 20.3% lung cancer, 8.3% colon cancer, 8.3% ovarian cancer, 4.3% prostate cancer.

3.1. Validity Results of Scales

The goodness-of fit index values obtained from the confirmatory factor baseline analyses of both scales were not acceptable to confirm the factor structure. For this reason, in accordance with the modification suggestions of the analyses, items 6-7 of CLS and items 1-4 and 3-5 of CRNSES were modified (Figure 1).

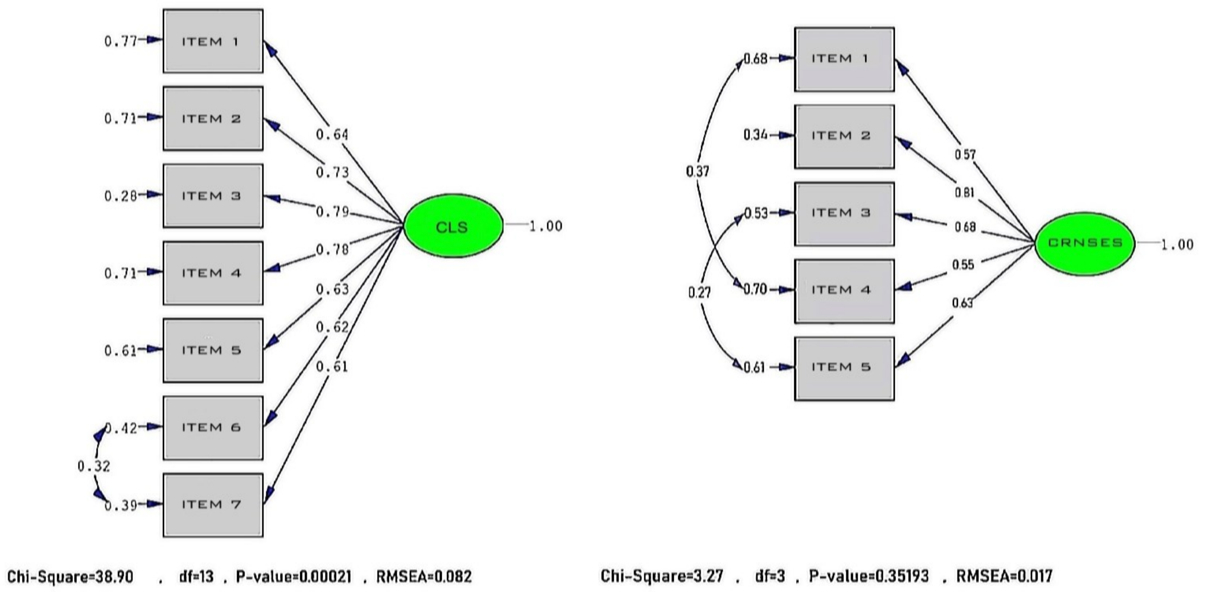


Figure 1. *CLS and CRNSES Modified PATH diagrams.*

After the modifications, the factor loads of the CLS were found to be between 0.61-0.78 and those of CRNSES were between 0.55-0.81. Table 1 shows the fit indexes of the modified models and basic models.

Table 1. *Model Fit Indexes of Basic Model and Post-Modification Scales (n=300)*

Cancer Loneliness Scale							
	X^2 /df	CFI	GFI	IFI	AGFI	RMSEA	Result
Basic Model	18.2	0.88	0.80	0.88	0.61	0.240	No fit
Modified Model	2,99	0,98	0,96	0.98	0.92	0.08	Perfect fit
Cancer-related Negative Social Expectations Scale							
	X^2 /df	CFI	GFI	IFI	AGFI	RMSEA	Result
Basic Model	27.13	0.83	0.85	0.41	0.54	0.296	no fit
Modified Model	1.33	1.00	1.00	1.00	0.98	0.02	Perfect fit

CFI: Comparative Fit Index; GFI: Goodness of Fit Index; IFI: İncremental Fit Index; AGFI: Adjusted Goodness of Fit Index; RMSEA: Root Mean Square Error Of Approximation.

UCLA General Loneliness scale was used to evaluate concurrent validity. The correlation between the scales was evaluated (Table 2).

Table 2. *Cancer Loneliness Scale, Cancer-related Negative Social Expectations Scale and UCLA Loneliness Scale Correlation*

	CLS	CRNSES	UCLA
Cancer Loneliness Scale (CLS)	*1*		
Cancer-related Negative Social Expectations Scale (CRNSES)	$r = 0.48$ $p = 0.000$	1	
UCLA Loneliness Scale	$r = 0.69$ $p = 0.000$	$r = 0.39$ $p = 0.000$	1

While a high positive correlation was found between CLS and UCLA General Loneliness scale, a moderate correlation was found between CRNSES and UCLA General Loneliness scale ($p < 0.001$). There was a statistically significant positive moderate correlation between CLS and CRNSES ($p < 0.001$) (Table 2).

3.2. Reliability Results of Scales

Item analysis of CLS revealed that the general average of the items was 2.6. The mean variation analysis was 0.83 (min. 2.05 - max. 2.88) (Hotelling's T-Squared = 223.25, *F* = 36.58, *p* = 0.000). When the total correlations were examined, the scale a was assumed as moderate and there were strong values between 0.52 and 0.70. Item analysis of CRNSES revealed that the general average of the items was 3.9. The mean variation analysis was 2.03 (min. 2.89 - max. 4.92) (Hotelling's T-Squared = 447.9, F = 110.87, *p* = 0.000). When the total correlations of the scale were investigated, it was found that the scale had moderate values between .60 and .62 (Table 3). The normal distribution of the scores obtained from the scale was evaluated by Skewness and Kurtosis coefficients. The Skewness and Kurtosis coefficients of CLS were -1.71 and -1.82, and those of CRNSES were -1.78 and -1.67, respectively.

Table 3. *Item analysis of Cancer Loneliness and Cancer-related Negative Social Expectations Scales*

Items	Mean	Standart deviation	Item-Total Correlation	Cronbach's Alpha Value of the Scale
Cancer Loneliness Scale *				
Item 1	2.74	1.08	0.519	0.864
Item 2	2.51	1.11	0.598	0.854
Item 3	2.67	0.94	0.740	0.834
Item 4	2.57	1.14	0.612	0.852
Item 5	2.05	1.00	0.577	0.857
Item 6	2.88	0.89	0.706	0.837
Item 7	2.82	0.87	0.709	0.836
Cancer-related Negative Social Expectations Scale **				
Item 1	2.89	1.78	0.621	0.793
Item 2	3.93	1.61	0.620	0.784
Item 3	4.60	1.37	0.613	0.781
Item 4	3.14	1.75	0.610	0.797
Item 5	4.92	1.24	0.605	0.785

** Hotelling's T-Squared 223,2   F=36.58   p=0,000*
*** Hotelling's T-Squared 447   F=110.9   p=0,000*

The mean score of CLS was 18.28 ± 5.2, the Cronbach's alpha coefficient was 0.88 and the Spearman-Brown correlation value was r = 0.81. The mean score of CRNSES was 19.5 ± 5.9, the Cronbach's alpha coefficient was 0.82 and the Spearman-Brown correlation value was *r* = 0.86 (*p* < 0.001) (Table 4).

Table 4. *Skewness-Kurtosis Coefficients and Internal Consistency Values of Scales (n=300)*

Scales	Mean±SS	Skewness	Kurtuosis	Cronbach Alpha	Spearman-Brown Correlation Coefficient	Guttman Split-Half
CLS	18.28±5.2	0.24±0.14 (-1.71)	-0.51±0.28 (-1.82)	0.88	0.81	0.78
CRNSES	19.5±5.9	-0.25±0.14 (-1.78)	-0.47±0.28 (-1.67)	0.82	0.86	0.82

CLS: Cancer Loneliness Scale
CRNSES: Cancer-related Negative Social Expectations

4. DISCUSSION and CONCLUSION

4.1.Validity of the Scales

Language adaptation of the scales was made according to WHO (2017) and ITC (2018) guidelines. In scope validity, expert opinions were evaluated with Davis technique and CVI was between 0.88 and 1.00. CVI value is expected to be greater than 0.80 (Davis, 1992; Erdoğan et al., 2017). According to our results, there is a consensus among the experts and the scales meet the criteria of scope validity.

If the scale in the study is newly developed, only Exploratory Factor Analysis (EFA) should be performed. However, if an existing scale is being adapted into another language, CFA should be performed (Erdoğan et al., 2017; Seçer, 2017). Within the scope of the CFA, direct and indirect effects between variables are tested in the context of a model constructed by researchers. Multiple indexes of fit are obtained in CFA and multiple indexes are evaluated together to assess whether the model is validated (Çokluk, Şekercioğlu & Büyüköztürk, 2014). Chi-square (X^2) value, X^2/SD value, Root Mean Square Error of Approximation (RMSEA), İncremental Fit Index (IFI), Comparative Fit Index (CFI), Goodness of Fit Index (GFI), and Adjusted Goodness of Fit Index (AGFI) were investigated in order to assess the model fit in the study. Chi-square is called a poor fit index, and high values indicate poor fit. The X^2/SD value can be used as a criterion for fit in large samples. Values of three and below are accepted as perfect fit (Çokluk et al., 2014). By analyzing the RMSEA value given under the path diagram, the difference between population and sample covariance is evaluated and this is expected to be between 0-1 (Çokluk et al., 2014; Seçer, 2017). For acceptable fit, IFI, CFI, GFI and AGFI values should be above 0.90, 0.95 and 0.85, whereas for perfect fit these values should be above 0.95, 0.97 and 0.90 (Seçer, 2017; Erdoğan et al., 2017; Seçer, 2018).

In the first analysis of the CLS (base model), the majority of fit indices were not acceptable (IFI 0.88, CFI 0.88, GFI 0.80, AGFI 0.61, X^2 /SD value18.2, RMSEA 0.240) (Çokluk et al., 2014; Marcoulides & Schumacker, 2014; Erdoğan et al., 2017; Seçer, 2018). In a CFA model, it may be difficult to redefine the model if the acceptance levels of the fit indices are not met. In this case, it is useful to examine the proposed modification suggestions given in analysis results (Çokluk et al., 2014). IFI 0.98, CFI 0.98, GFI 0.96, AGFI 0.92, and X^2 /SD 2.99 were obtained in the post-modification model of the CLS, and perfect fit values were obtained. RMSEA was 0.08. Adams et al. also reported that in the final model, one-dimensional CLS showed perfect fit (RMSEA = 0.03; CFI = 1.00; X^2 (13) = 15.73, P = 0.26) (Adams et al., 2017).

In the base model of the CRNSES, the fit indices were not acceptable (IFI 0.41, CFI 0.83, GFI 0.85, AGFI 0.54, X^2/SD 27.13, RMSEA 0.296) (Çokluk et al., 2014; Marcoulides & Schumacker, 2014; Erdoğan et al., 2017; Seçer, 2018). In the post-modification model, all indices showed perfect fit (X^2/SD 1.33, GFI 1.00, AGFI 1.00, CFI 1.00, RMSEA 0.02). The single-factor structure of CRNSES consisting of 5 items was confirmed as a model. In the study in which CRNSES was developed, it was stated that perfect fit was obtained with the final model (RMSEA = 0.03; CFI = 1.00; X^2 (4) = 4.70, P = 0.32) (Adams et al., 2017). According to the results of our study, the structure of CLS and CRNSES was supported by confirmatory factor analysis.

A positive, strong and significant correlation was found (r = 0.69) between CLS and UCLA General Loneliness Scale, while a positive moderate correlation was found between CRNSES and UCLA General Loneliness Scale. It can be said that the scales are valid for measuring the loneliness level as well as negative social expectations of cancer patients. Adams et al. also noted a strong correlation between CLS and UCLA (r = 0.67), and between CRNSES and UCLA (r = 0.47) in a positive manner (Adams et al., 2017). The correlation between CLS and general loneliness scale obtained in our study shows that CLS is a valid scale.

There was a positive moderate correlation between CLS and CRNSES (*r* = 0.48). In the original scale, there was a strong positive correlation between CLS and CRNSES (*r* = 70) and it was reported that findings consistent with loneliness theory were obtained (Adams et al., 2017). The correlation between CLS and CRNSES is important in terms of focusing on cancer-specific experiences.

4.2. Reliability of Scales

Whether the study data fits normal distribution is important for the reliability and generalizability of the research results and it can be evaluated by performing different normality tests. The distribution is considered normal if the resulting value is between -1.96 and +1.96 when the Skewness-Kurtosis coefficients are divided by standard errors (Can, 2018). In our study, it was observed that both scales were within this range and showed normal distribution (CLS: -1.71 and -1.82, CRNSES: -1.78 and -1.67).

Item analysis was carried out in order or identify the discriminative power of the scales (Seçer, 2017). As a result of item analysis, item-total correlations of CLS was found to be between 0.52 and 0.70. Item-total correlations of CRNSES was found to be ranging from 0.60 to 0.62. Items with a value of 0.30 and above are considered to have good discriminative power in terms of the measured property (Seçer, 2017). Item-total correlations of the scales were sufficient. It can be said that the item averages in both scales are different from each other, the items are not perceived by the participants with the same approach, the difficulty levels and measurement abilities of the items are different, and each item should be present in the scales (p <0.001). After the analysis (CLS: Hotelling's T^2 test = 223.2, F = 53.44, $p < 0.001$; CRNSES: Hotelling's T^2 test = 447.9, F = 53.44, $p < 0.001$), it was found that the nurses did not perceive the items with the same approach, and answered the items by directly reflecting their opinions at different degrees. The consistency of the items constituting a test among each other indicates internal consistency. Cronbach's alpha method is one of the most frequently used methods for determining internal consistency in scale adaptation studies (Seçer, 2017; Erdoğan et al., 2017; Can, 2018). Evaluation of Cronbach's alpha coefficient is as follows: 0.40–0.60 low reliability, 0.60–0.80 moderate, and 0.80-1.00 high reliability (Tavşancıl, 2014). The internal consistency coefficient of both scales was above 0.80 and the scales were found to be highly reliable. According to these results, it can be said that the items of the scales are consistent with each other and the scales are homogeneous. Adams et al. determined the Cronbach's alpha coefficient of CLS as 0.94. CRNSES was found as 0.90. The internal consistency coefficients of the scales are highly reliable.

In our study, the mean score of CLS was 18.28 ± 5.2 (min.7- max.32), and the mean score of CRNSES was 19.5 ± 5.9 (min.5 - max.30). These results showed that the patients included in the study had moderate cancer-related loneliness but their negative social expectations were above the moderate level. Negative social expectations may be associated with cancer experience in particular. For example, their friends and family may show major level of support and sympathy after the diagnosis. The patients may feel, disappointed if such behaviors are not seen. Loneliness theory and studies have shown loneliness could have original precipitates in cancer patients (Adams, 2016).

Split-half reliability test have been developed to eliminate the time problem that emerges in the test-retest method and the difficulty of finding equivalent forms in the validity of equivalent forms (Seçer, 2017). If the correlation coefficient between the split-half of the scale is 0.70 or above, its internal consistency is high (Boyle, Saklofske, & Matthews, 2015; Erdoğan et al., 2017). Spearman-Brown correlation value was r = 0.81 for CLS and r = 0.86 for CRNSES ($p <$ 0.001). When CLS and CRNSES are evaluated as a whole, it can be said that they consist of closely related items and their internal consistency is high.

In conclusion, CLS and CRNSES are valid and reliable scales that can be used in Turkish society. These scales will help in the assessment and identification of loneliness and negative social expectations, which is a deficiency in treatment and care practices in cancer patients. In this context, the development of loneliness decreasing interventions can be crucial in terms of making the mental and physical health conditions of cancer patients better. In addition, reducing the disease-related mortality and morbidity with the psychosocial support given to the patients will increase the life standards of the family members and patients and will provide further benefit in terms of public health. The scales can be used in clinical practice and on cancer patients in the field, and also in academic studies that will contribute to the literature. The fact that CLS is shorter compared to the current loneliness scales is regarded as an advantage in terms of convenient and faster response by cancer patients.

Acknowledgements

This study is a project financed by Pamukkale University Scientific Research Projects (Project no: 2018 SABE020).

Declaration of Conflicting Interests and Ethics

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

ORCID

Ebru KARA https://orcid.org/0000-0003-0326-8734
İlgün ÖZEN ÇINAR https://orcid.org/0000-0001-5774-5108

5. REFERENCES

Adams, R.N. (2016). Measures of Cancer-Related Loneliness and Negative Social Expectations: Development and Preliminary Validation, in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy (Doctoral dissertation). Indiana, Purdue University, 2016.

Adams, R.N., Mosher, C.E., Rand, K.L., Hirsh, A.T., Monahan, P.O., Abonour, R. & Kroenke, K. (2017). The Cancer Loneliness Scale and Cancer-Related Negative Social Expectations Scale: Development and validation. *Qual Life Res.*, *26*(7), 1901-1913. https://doi.org/10.1007/s11136-017-1518-4

Adams, R.N., Mosher, C.E., Winger, J.G., Abonour, R., & Kroenke, K. (2018). Cancer-related loneliness mediates the relationships between social constraints and symptoms among cancer patients, *J Behav Med.*, *41*(2), 243–252. https://doi.org/10.1007/s10865-017-9892-5

Boyle, G. J., Saklofske, D. H., & Matthews, G. (2015). Criteria for Selection and Evaluation of Scales and Measures. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of Personality and Social Psychological Constructs* (p. 3–15). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-386915-9.00001-2

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality World wide for 36 cancers in 185 countries. *CA Cancer J Clin., 68*(6), 394-424. https://doi.org/10.3322/caac.21492

Brintzenhofe-Szoc, K.M., Levin, T.T., Li, Y., Kissane, D.W., & Zabora, J.R. (2009). Mixed anxiety/depression symptoms in a large cancer cohort: Prevalence by cancer type. *Psychosomatics, 50*(4), 383-391. https://doi.org/10.1176/appi.psy.50.4.383

Can, A. (2018). *SPSS ile Bilimsel Araştırma Sürecinde Nicel Veri Analizi*, Bir Ölçme Aracı ile Yapılan Ölçümün Güvenirliliğini Belirleme (p. 385-394). Ankara: Pegem Yayıncılık.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (Eds). (2014). *Sosyal Bilimler için Çok Değişkenli İstatistik SPSS ve LISREL Uygulamaları*, Yapısal Eşitlik Modeli (p 251-407). Ankara: Pegem Yayıncılık.

Davis, L. (1992). Instrument Review: Getting the Most From a Panel of Experts, Clinical Methods. *Appl Nurs Res, 5*(4), 194-197. https://doi.org/10.1016/S0897-1897(05)80008-4

Deckx, L., Akker, M., & Buntinx, F. (2014). Risk factors for loneliness in patients with cancer: A systematic literature review and meta-analysis. *Eur. J. Oncol. Nurs.*, *18*(5), 466-477. https://doi.org/10.1016/j.ejon.2014.05.002

Demir, A. (1988). UCLA Yalnızlık Ölçeğinin Geçerlik ve Güvenirliği [Validity and Reliability of UCLA Loneliness Scale] *Türk Psikoloji Dergisi*, *6*(23),14-18.

Dodds, S.E., Pace, T.W., Bell, M.L., Fiero, M., Negi, L.T., Raison, C.L., & Weihs, K.L. (2015). Feasibility of Cognitively-Based Compassion Training (CBCT) for breast cancer survivors: a randomized, wait list controlled pilot study. *Support Care Cancer, 23*(12), 3599-3608. https://doi.org/10.1007/s00520-015-2888-1

Erdoğan, S., Nahcivan, N., & Esin, M. (Eds) (2017). *Hemşirelikte Araştırma, Süreç Uygulama ve Kritik*, Veri Toplama Yöntem ve Araçları &Veri Toplama araçlarının Güvenirlik ve Geçerliliği (p. 195-232). İstanbul: Nobel Tıp Kitapevleri

Fanakidou, I., Zyga, S., Alikari, V., Tsironi, M., Stathoulis, J., & Theofilou, P. (2018). Mental health, loneliness, and illness perception out comes in quality of life among young breast cancer patient safter mastectomy: the role of breast reconstruction. *Qual Life Res,, 27*(2), 539–543. https://doi.org/10.1007/s11136-017-1735-x

International Test Commission (ITC) (2018) Guidelines for translating and adaptation tests. International Journal of testing, 18:101-134. Retrevied July 31, 2019, from http://dx.doi.org/10.1080/15305058.2017.1398166

Jaremka, L.M., Andridge, R.R., Fagundes, C.P., Alfano, C.M., Povoski, S.P., Lipari, A.M. Kiecolt-Glaser, J.K. (2014a). Pain, depression, and fatigue: Loneliness as a longitudinal risk factor. *Health Psychol*, *33*(9), 948-957. https://doi.org/10.1037/a0034012

Jaremka, L.M., Fagundes, C.P., Glaser, R., Bennet, J.M., Malarkey, W.B., & Kiecolt-Glaser, J.K. (2013). Loneliness Predicts pain, depression, and fatigue: Understanding the role of immune dysregulation. *Psychoneuroendocrinology*, *38*(8), 1310-1317. https://doi.org/10.1016/j.psyneuen.2012.11.016

Jaremka, L.M., Peng, J., Bornstein, R., Alfano, C.M., Andridge, R.R., Povoski, S.P. …Kiecolt-Glaser, J.K. (2014b). Cognitive problems among breast cancer survivors: Loneliness enhances risk. *Psycho-Oncology, 23*(12), 1356-1364. https://doi.org/10.1002/pon.3544

Kara, E. (2019). *Kanser Yanlızlık Ölçeği ve Kansere İlişkin Negatif Soyal Beklentiler Ölçeğinin Türkçe Geçerlilik ve Güvenirliği*. Pamukkale Üniversitesi Sağlık Bilimleri Enstitüsü, Türkiye, Yüksek Lisans Tezi.

Kroenke, K., Johns, S.A., Theobald, D., Wu, J., & Tu, W. (2013). Somatic symptoms in cancer patient straje ctoryover 12 months and impact on functional status and disability. *Supportive Care in Cancer (MASCC)*, *21*(3), 765–773. https://doi.org/10.1007/s00520-012-1578-5

Macmillan Cancer Support. Jones C. (2014). Lonely cancer patients three times more likely to struggle with treatment. http://www.macmillan.org.uk

Marcoulides, G.A., & Schumacker, R.E. (Ed.). (2014). New Developments and Techniques in Structural Equation Modeling. New Developments and Techniques in Structural Equation Modeling. Psychology Press. https://doi.org/10.4324/9781410601858

Pehlivan, S., Ovayolu, O., Ovayolu, N., Sevinç, A., & Camcı, C. (2012). Relationship between hopelessness, loneliness, and perceived social support from family in Turkish patients with cancer. *Support Care Cancer*, *20*(4), 733-739. https://doi.org/10.1007/s00520-011-1137-5

Russell, D.W. (1996). UCLA Loneliness Scale (Version 3): Reliability, validity, and factor structure. *J Pers Assess*, *66*(1), 20-40. https://doi.org/10.1207/s15327752jpa6601_2

Seçer, İ. (2017). *SPSS ve LISREL ile Pratik Veri Analizi,* (p. 211-222). Ankara: Anı Yayıncılık.

Seçer, İ. (2018). *Psikolojik Test Geliştirme ve Uyarlama Süreci*, (p.18-104). Ankara: Anı Yayıncılık.

Streiner, D.L., & Kottner, J. (2014). Recommendations for Reporting the Results of Studies of Instrument and Scale Development and Testing. *J. Adv. Nurs.*, *70*(9), 1970–1979. https://doi.org/10.1111/jan.12402

Tabrizi, F.M., Radfar, M., & Taei, Z. (2016). Effects of supportive-expressive discussion groups on loneliness, hope and quality of life in breast cancer survivors: a randomized control trial. *Psycho-Oncology*, *25*(9), 057-1063. https://doi.org/10.1002/pon.4169

Tavşancıl, E. (2014). *Tutumların Ölçülmesi ve SPSS ile Veri Analizi*, Ölçme ile ilgili Temel Kavramlar (p.3-58). 5. Baskı, Ankara: Nobel Yayın Dağıtım.

Wells, M., & Kelly, D. (2008). The loneliness of cancer. *Eur. J. Oncol. Nurs.*, *12*(5), 410– 411. https://doi.org/10.1016/j.ejon.2008.11.003

World Health Organization (WHO) (2017). Process of translation and adaptation of instruments. Retrieved July 10, 2019, from http://www.who.Int/subtance_abuse/research_tools/translation/en/

World Health Organization. *Early Detection of Cancer*. World Health Organization; 2019. Retrieved July 31, 2019, from https://www.who.int/cancer/detection/en/

6. APPENDIX: Turkish Form of The Cancer Loneliness Scales (Sample Item)

Table A1. *A few sample items from the Cancer Loneliness Scale (Kara, 2019).*

Aşağıdaki ifadeler, kanser teşhisi konulduktan sonra insanların nasıl hissettiğini açıklar. Her ifade için, boşluklara ne sıklıkta o şekilde hissettiğinizi yazın.

ASLA	NADİREN	BAZEN	SIKLIKLA	HER ZAMAN
1	2	3	4	5

1.Kanser teşhisi konulduktan sonra ne sıklıkta, en yakın arkadaşlarının ya da aile bireylerinin seni yanlış anladığını hissediyorsun? …………

2.Kanserle mücadelende, ne sıklıkta diğer insanların sana yeterince destek olamadıklarını düşünüyorsun? ………..

3.Kanser teşhisi konduktan sonra ne sıklıkta çevrenizdeki insanlarla çok fazla ortak noktanız olmadığını hissediyorsun? ………..

4.…………………………………………………………………………………………….. ………..

5. ……………………………………………………………………………………………. …………

6. …………………………………………………………………………………………….. …………

7.……………………………………………………………………………………………... …………

Table A2. *A few sample items from the Cancer-related Negative Social Expectations (Kara, 2019).*

Lütfen her bir satırda tek bir kutuyu işaretleyerek soruları cevaplayınız.

	Kesinlikle Katılmıyorum	Orta Seviye Katılmıyorum	Biraz Katılmıyorum	Biraz Katılıyorum	Orta Seviye Katılıyorum	Kesinlikle Katılıyorum
1.Eğer insanlara kanser geçmişimden bahsedersem endişelenir ve benim yanımda rahat davranamazlar…	□	□	□	□	□	□
2.Eğer insanlar, kanser hastalığım hakkında konuşmak istemezse bunu duymak istemediklerini düşünürüm…	□	□	□	□	□	□
3………………………… ………………………. ………………………… ……	□	□	□	□	□	□
4………………………… ……………………… …………………….	□	□	□	□	□	□
5………………………… ………………………. ………………………… ………………	□	□	□	□	□	□

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 404–419
https://doi.org/10.21449/ijate.781366
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Review Article**

Establishing survey validity: A practical guide

William W. Cobern [1,*], **Betty AJ Adams** [2]

[1]The George G. Mallinson Institute for Science Education, Western Michigan University, Kalamazoo, MI, USA

ARTICLE HISTORY

Received: May 22, 2020

Accepted: Aug. 15, 2020

KEYWORDS

Pretesting,
Cognitive interviews,
Reliability,
Research Methods,
Survey Methods,
Validity

Abstract: What follows is a practical guide for establishing the validity of a survey for research purposes. The motivation for providing this guide is our observation that researchers, not necessarily being survey researchers per se, but wanting to use a survey method, lack a concise resource on validity. There is far more to know about surveys and survey construction than what this guide provides; and this guide should only be used as a starting point. However, for the needs of many researchers, this guide provides sufficient, basic information on survey validity. The guide, furthermore, includes references to important handbooks for researchers needing further information.

1. INTRODUCTION

We have written this practical guide because of a dispute that arose between two faculty members and a student. The faculty members criticized the student for having insufficiently established the validity of a survey she had created. As the student was working under our supervision, the criticism was surprising. On the other hand, we quickly realized that the situation constituted a proverbial "teachable moment." Even though the student had taken a course on survey development and we had discussed the methodology, we realized that neither students nor faculty had a practical guide on how to establish survey validity, or what that even means. This document is an attempt to fill that need.[†,‡,§] These are not survey researchers per se, but researchers who on occasion need to develop a survey for the purposes of their research interests.

CONTACT: William W. Cobern ✉ bill.cobern@wmich.edu 🖃 The George G. Mallinson Institute for Science Education, Western Michigan University, Kalamazoo, MI, USA

[†] This guide does not address the purposes for survey research. The assumption of this guide is that the researcher has already made the decision to use a survey. This guide is solely about the production of a valid survey for research purposes.

[‡] Boateng et al. (2018) offers a similar practical guide but from a different perspective with somewhat different coverage.

[§] Much of what is in this practical guide can also be applied to the development and validation of interview protocols.

ISSN-e: 2148-7456 /© IJATE 2020

At the start it is important to distinguish between surveys and tests, though in fact much of this practical guide is also relevant to test construction. Tests and surveys have much in common, indeed, sometimes it is difficult to tell the difference. For example, is the *Student Understanding of Science and Scientific Inquiry* (SUSSI) a test or a survey? Is the *Views of Nature of Science Questionnaire* (VNOS) a test or a survey? Is the *Measure of Acceptance of the Theory of Evolution* (MATE) a test or survey? Is the PEW instrument for assessing public knowledge of science a test of knowledge or a survey of knowledge? Could be either. For the purposes of this practical guide, we make the following distinction. Surveys (or questionnaires[**]) typically collect information about attitudes or opinions, can also be used to survey knowledge, but are typically not associated with instructional settings. On the other hand, tests are almost always about knowledge or skills and, unlike surveys, tests generally are associated with instruction. This is not a hard and fast distinction, however, so in this practical guide we will use examples that some people may think of as tests; it makes no difference to the procedures we present.

This practical guide is purposefully simple as the objective is to provide practical guidance on a few basic things that all researchers should observe for establishing survey validity. Furthermore, one can think of survey construction as serving one or two purposes. Researchers may construct survey instruments because they need an instrument to collect data with respect to their specific research interests. The survey is not the focus of the research but a tool, an artifact of conducting research. Other people may decide to use the researcher's instrument as they see fit, though it was not the researcher's intention to provide a new instrument for other researchers to use. For example, Barbara Greene studies cognitive engagement and for this purpose she and her colleagues have developed a number of survey-type instruments. She writes about getting regular requests from others wishing to use her cognitive engagement scales, which came as a surprise to her group as they developed the scales for their own research purposes (Greene, 2015). They were not in the business of developing instruments for general research use. On the other hand, some research is specifically about survey construction of which there are many examples including Lamb, Annetta, Meldrum, and Vallett (2012), Luo, Wang, Liu, and Zhou (2019), Staus, Lesseig, Lamb, Falk, and Dierking (2019).

Survey development can involve powerful statistical techniques such as Item Response Theory (Baker, 2001) or Rasch Modelling (Boone, 2016). One is more likely to see these techniques used when a survey is developed for broad use. These techniques are less common when a survey instrument is developed as an internal artifact for conducting specific research. Perhaps more often one will see researchers employ factor analyses as part of survey development. This practical guide does not address either Rasch Modelling or Item Response Theory, and only mentions factor analysis in passing. Our focus is on the development of narrowly focused surveys designed for the research a person wishes to pursue, and not on the development of a survey for others to use. Of course, for whatever reason someone produces a survey, as noted above, that survey is likely to get used by others regardless of the originator's intention for the survey.

Surveys serve a broad range of purposes. Some are simply seeking factual or demographic information. We may want to know the age range across a group of people. We may wish to ask students enrolled in a particular course what their majors are. We might be interested in how a group of people prioritizes a set of unambiguous entities. On the other hand, we might be interested in using surveys to gauge far more complex constructs such as attitudes, behaviors, or cognitive engagement. The latter are much more difficult to develop and validate than are the former.

[**] We do not think that there is anything in the literature that provides a strong rationale for distinguishing between surveys and questionnaires. For all practical purposes, there is no difference. The research literature, however, typically uses the word survey.

Whether using sophisticated methods such as Rasch Modelling, Item Response Theory, or factor analysis, or more basic methods, whether developing a simple survey or a rather complex one, every researcher begins with three questions that are not necessarily easy to answer:[††]

1) What is it that I want to learn from the responses on my instrument?
2) What assurance can I have that my respondents understand what I am asking?
3) How can I be reasonably sure that the responses my respondents give to my items will be the same responses they give to the same items two weeks later?

The first and second questions are about instrument validity, and the third question is about instrument reliability.

2. EVIDENCE SUPPORTING VALIDITY

What is this idea of validity? Here is an example to help illustrate the general idea of validity. If you give students a set of questions having to do with their interest in science and they consistently respond about their interests in the arts, there is a problem. The questions prompted consistent[‡‡] responses but the responses are not about the information you were seeking. Somehow, the questions give the respondents the wrong idea that you wanted to know about their interest in the arts when what you wanted was to know about their interest in the sciences. Your questions are not valid with respect to the information you are trying to get. A test item or survey item (and this applies to interview items as well) has validity if the reader of the item understands the item as intended by the item's creator. As stated in the 2018 *Palgrave Handbook of Survey Research* (Vannette & Krosnick, 2018):

> An important aspect of validity is that the survey is designed in such a way as to minimize respondent error. Respondent error has to do with the respondent responding to an item in some way that is different from the researcher's intention. (Krosnick, 2018, p. 95)

Validity is an evidence-based argument. The researcher provides evidence that the instrument is valid with respect to its intended purpose and audience. According to the 2014 *Standards for Educational and Psychological Testing*,

> Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use. (AERA, APA, NCME, 2014, p. 11)

At least since the 1999 *Standards* edition, measurement experts in education and psychology have ceased referring to distinct types of validity (e.g., content or construct validity)[§§], preferring to view validity as a unitary concept represented by the "degree to which all accumulated evidence supports the intended interpretation of test scores for the proposed use" (AERA, APA, NCME, 2014, p. 14). Moreover, as one might expect, there are various sources and types of evidence:

> That might be used in evaluating the validity of a proposed interpretation of test scores for a particular use. These sources of evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. (AERA, APA, NCME, 2014, p. 13-14)

[††] Our epistemological perspective is that survey development and validation are processes that need to proceed hand-in-hand. We do not consider it wise for the researcher to separate these processes into a sequence of development first followed by validation

[‡‡] Consistency has to do with reliability and is discussed later.

[§§] See Ruel et al. (2016) for an example from sociology of researchers retaining the old system.

Furthermore, “the wide variety of tests and circumstances makes it natural that some types of evidence will be especially critical in a given case, whereas other types will be less useful” (AERA, APA, NCME, 2014, p. 12).

It is beyond the scope of this practical guide to present much detail on the various types of evidence that can be used in support of validity. For that purpose, readers should consult authoritative documents such as the 2014 *Standards for Educational and Psychological Testing* or the 2018 *Palgrave Handbook of Survey Research*. However, for practical purposes, there are two areas of importance for establishing evidence of validity: a validated model that provides the basis for an instrument, and the items composing an instrument.

2.1. Foundational Model

A valid survey requires a theoretical model of what it is the researcher wants to find out by having people respond to survey items. The foundational model answers the question: *What is it that I want to learn from the responses on my instrument*? Answering this question involves obtaining or building a validated, theoretical model for what the researcher wants to know. Beware of the temptation just to write items straightaway. This happens far too many times where the researcher completely skips the idea of theoretical model building and jumps directly into writing items (or questions).[***] These are items simply coming to one’s mind but lacking theoretical foundation. Such items are *ad hoc*, and an instrument built on *ad hoc* items is not a research-worthy instrument. There is already a validity issue because there is no foundation for the survey. The first line of validation evidence for survey items is the foundational model.

While there probably are many ways to develop a foundational model, these ways certainly include theory-driven model development, statistically-derived model development, and grounded theory model development. Theory-driven model development is a top-down approach in contrast to the bottom-up approach of statistically-derived model development and grounded theory model development. Bottom-up model development is essential when the researcher has no *a priori* model or theory on which to build a survey. In that situation, the model has to be built inductively from data collected from the type of people who would ultimately become subjects of research where the survey is used, or possibly built inductively from expert opinion. Bottom-up model development oftentimes involves a combination of grounded theory and statistical analysis. For example, let’s say you are interested in the goals that college faculty have for chemistry lab instruction and you would like to survey a large number of college chemistry faculty to determine what goals are most frequent. Bruck and Towns (2013) developed such a survey that began with a grounded theory approach. Initially, the researchers collected qualitative data from interviews with college chemistry faculty on the goals they had for chemistry lab instruction (Bruck, Towns, & Bretz, 2010). Subsequently,

> An initial pool of survey items was developed from findings of the qualitative study. Questions constructed from key interview themes asked respondents to identify the frequency of certain laboratory practices, such as conducting error analyses or writing formal laboratory reports. (Bruck & Towns, 2013, p. 686)

When these researchers say that they developed an initial pool of items drawing from the findings of their qualitative study, they are essentially describing a grounded theory approach. They are “on the ground” with college chemistry faculty finding out directly from them what their goals are. However, this data has no structure; it represents no model. To create a foundational model that provides structure for a survey based on the ideas coming directly from the faculty, the researchers turned to statistical methods. The researchers drafted a survey using

[***] People use the terms ‘item’ and ‘question’ interchangeably with regard to surveys. ‘Item’ is the more general term but items on a survey are all questions in that each item represents a request for information whether it is, for example, one's birthday or one's opinion registered on a Likert scale.

these items that they then distributed to a large number of chemistry faculty. They subjected the resulting data to statistical procedures (correlation tables, Cronbach’s α, Kaiser-Meyer-Olkin tests, and factor analysis) resulting in a seven-factor model:

Research Experience	Transferable Skills (Lab-Specific)
Group Work and Broader Communication Skills	Transferable Skills (Not Lab-Specific)
Error Analysis, Data Collection and Analysis	Laboratory Writing
Connection between Lab and Lecture	

In the process, the researchers dropped items not fitting this model (i.e., those having low statistical value) resulting in the 29-item *Faculty Goals for Undergraduate Chemistry Laboratory Survey*, a survey for which the foundational model was derived bottom-up using a combination of grounded theory and statistical methods. The validity lines of evidence include the initial qualitative data gathered from interviews and the subsequent statistical analyses of data. For an instrument derived from a combination of grounded theory and statistical methodology, the building and validation of the model and the instrument are intertwined. They go hand-in-hand.

The development of a theoretically derived foundational model is much different, though the question remains the same: *What is it that I want to learn from the responses on my instrument*? The difference is that the researcher already has a model or theory on which to base the instrument; hence, the development approach is top-down. The survey is derived deductively from the model. Such models can come from the literature (which is often the case) or researchers construct the model by drawing from the literature. In either case, the connection to the literature validates the model. Moreover, it is possible for researchers to invent a model to suit their philosophical positions and research interests. Our first example comes from research conducted by the first author and is an example of a model drawn from the literature (Cobern, 2000).

Cobern, Gibson, and Underwood (1999) and Cobern (2000) reported investigations of how students conceptualize nature, that is, the natural world. The studies had to do with the extent to which students voluntarily introduce science into their explanations about nature. These were interview studies rather than survey studies; but the theoretical modeling would have been the same had Cobern decided to collect data using a survey. A wide-ranging review of the literature led to a model involving four categories of description along with a set of disparate adjectives that could be used to represent each category description (see Table 1).

This model represents what Cobern wanted to learn from the study. He wanted to learn the various ways in which students might describe nature, and for reasons described in the published papers, he based the interview protocol on this *a priori*, theoretical model. Basing the interview protocol on the theoretical model provides the first line of validity evidence. The same would be true if he had decided to use a survey method. Deriving the survey from a literature-validated model[†††] provides the first line of validity evidence for the survey.

[†††] The literature-based validation of a model does not mean that one particular model is the only one a researcher could validate from literature. Undoubtedly, in most situations, literature can validate a number of different models. Therefore, the onus is on researchers to explain why they built a particular model and on readers to judge that explanation.

Table 1. *Modeling: what is nature? (Cobern, 2000, p. 22)*

Epistemological Description: (Reference to knowing about the natural world.)	confusing mysterious	unexplainable unpredictable	understandable predictable knowable
Ontological Description: (Reference to what the natural world is like.)	material matter living complex orderly beautiful	dangerous chaotic diverse powerful changeable	holy sacred spiritual unchangeable pure
Emotional Description: (Reference to how one feels about the natural world.)	peaceful	frightening	"just there"
Status Description: (Reference to what the natural world is like now.)	"full of resources" endangered	exploited polluted	doomed restorable

It is important to understand that the above examples involve categories that subsume items or interview questions. Respondents address the items, not the categories. For example, the Bruck and Towns (2013) survey does not explicitly ask respondents about "research experience," which is one of their categories. "Research experience" is too ambiguous a term (see section on item clarity below) to ask about it explicitly. Rather, respondents see a set of clearly stated items that according to the researchers' model represents "Research Experience." Thus, respondents do not need to understand the construct; they only need to understand the language of the items in which the construct is expressed. A consequence of such modeling is that the internal consistency of categories needs to be checked every time the instrument is used. Researchers should not assume "once validated, always validated."

The Cobern (2000) model was constructed *from* the literature; however, in other cases, a top-down model may be found *directly* in the literature. In other words, the model is not derived from the literature but is literally borrowed from the literature. For example, Haryani, Cobern, and Pleasants (2019) investigated Indonesian teachers prioritizing of selected curriculum objectives. Their national Ministry of Education establishes the Indonesian curriculum and it is incumbent upon all Indonesian teachers to know and follow this official curriculum. Haryani et al. (2019) was specifically interested in the new addition of 21st Century Learning Skill objectives to the curriculum (creativity and innovation, critical thinking, problem-solving, collaboration, and communication skills), and how teachers prioritized these new objectives. The model for the research survey (Table 2 below) came directly from the official curriculum. Basing the survey items on this theoretical model read from the literature (i.e., the official curriculum) provided the first line of validity evidence.

Summarizing this section, establishing the validity of an instrument begins with clearly answering this question: what is it that I want to learn from the responses on my instrument? Answering this question begins with having a validated, theoretical model (a foundational model) for what the researcher wants to know. The next section is about constructing a survey based on a model: item fit, instrument length, item format, item discrimination, item clarity, order of items, and item effectiveness.

Table 2. *Modeling: teacher C13-curriculum priorities*

The C13 Curriculum Content	Outcomes
Traditional C13 content	Science Content
Recent C13 additions	Science Processes
21st Century Learning Skill	Creativity and Innovation
21st Century Learning Skill	Critical Thinking
21st Century Learning Skill	Problem Solving
21st Century Learning Skill	Collaboration
21st Century Learning Skill	Communication Skills
C13 Irrelevant content	History of Science
C13 Irrelevant content	Writing Skills
Participant demographics[‡‡‡]	Gender, school type

2.2. Fitting Items to The Model

As noted earlier, sometimes the researcher is tempted to start instrument development by simply writing items as they come to mind. That temptation needs to be avoided by giving due attention to first building a model or acquiring one. With a model in hand to inform the development of the instrument, the researcher can either write original items or find useful items in the literature to use as-is or revised, or build an instrument from a combination of both. As items are gathered, they need to be fitted to the model. The model serves as a device for disciplining the selection of items. Furthermore, the fit should be validated by persons external to the instrument development process. In other words, the researcher should have a few, knowledgeable people check items for fit with the model.

Instrument length: Selecting items (or writing items) raises questions about the number of items, the wording of items, and item type. Regarding the number of items and thus the length of a survey, the rule of thumb is that shorter is better than longer. As noted by Krosnick (2018, p. 95), "the literature suggests that what goes quickly and easily for respondents also produces the most accurate data." In other words, the threat to validity increases with instrument length.

Researchers need to minimize the length of a survey; but if a survey has to be long then precautions are needed because excessive length will very likely introduce response errors.[§§§] For example, Nyutu, Cobern, and Pleasants (2020) needed student responses to 50 items in order to build a model (using a bottom-up approach) for their work on faculty goals for laboratory instruction. The researchers were concerned that students would not take the last items seriously given the length of the survey. To mitigate the potential problem, the researchers used five different forms of the survey where the item order was different on each form. By doing this, response errors in the last items would not be concentrated in the same items. This approach does not eliminate the problem but it at least eliminates the impact on specific items. Another approach would have been to use filter items toward the end of the survey. The researchers could have added one or two items toward the end that requested a specific response. For example, an item could have simply read, "For this item, select 3." Thus, any survey that did not have a "3" for this item would have to be considered suspect. There are no perfect solutions when working long surveys but there are strategies, each with its advantages and disadvantages.

[‡‡‡] The inclusion of last three elements in this model, which are not 21st Century Learning Skills, is explained later.

[§§§] For example, if a survey is very long then respondents may not pay attention to the last items because they have become tired of responding to so many items.

Of course, the best thing to do is to keep a survey short, and the model will help limit the number of items selected. However, researchers oftentimes want demographic data and this is where survey length can get out of hand. Note that the last entry in Table 2 is participant demographics. The researchers specifically placed demographics in the model as a reminder to only ask for demographics that were important with respect to the rest of the model. For example, if the researcher does not have a good reason (that is, reasons relevant to teacher prioritizing of curriculum objectives) for asking teachers about their age, then the researcher should not ask for age. The researcher should only ask for demographics that are important to the study or for which the researcher has good reason to think could be important. Researcher discipline about demographic information helps keep survey length reasonable, bearing in mind that excessive survey length poses a threat to validity.

2.3. Item Format

The type of items to be used is another important question specific to survey development. Survey items frequently use Likert scales, which raises the question of how many points should be on a scale. Conventional wisdom is to use an odd number such as five or seven (Krosnick, 2018, p. 99). However, sometimes a researcher wants to avoid having respondents select a middle or "neutral" position, in which case the scale has to be an even number. Too few points or too many points threaten validity, and could either blur or exaggerate variation.

Survey items are oftentimes about information where the Likert format is not useful. Writing such items is fairly straightforward when the information is simple such as age. Asking how often somebody engages in activity can be trickier. For example, asking how often students watch YouTube videos has to begin with the assumption that students are unlikely to have a good idea of exactly how much time they spend per week watching YouTube videos. Hence, asking how many hours some spend watching YouTube each week is likely to return unreliable responses, mere guesses. Students will be more reliable approximating their viewing time given a choice of time intervals such as a) 0 to 5 hours per week, 6 to 10 hours per week etc. The challenge for the researcher is to create reasonable time intervals. While there are no guidelines or rules to help the researcher, the researcher can check the literature to see the kind of time intervals that have been used by other researchers and use that as a guide; or the researcher can create the intervals with respect to the needs of the research. By the latter we mean that the researcher decides reasonable magnitudes for the poles based on the nature of the research questions. Again, using YouTube viewing as an example, the researcher may decide that watching YouTube 10 hours a week would be a lot and that few students are likely to do that. On the other hand, the researcher might reason that most students would watch for at least an hour. Following this line of reasoning, the lower time interval might be 0 to 1 hour with the upper interval being 10 hours or more: a) 0-1 hrs, b) 2 to 5 hrs c) 6 to 10 hrs d) 10+ hrs. And as should be common practice, it is a good idea to have somebody outside of the research check the researcher's decision. For example, if the item is intended for students then the researcher should ask a few students about the item. For example, the researcher might ask the students if these are the time intervals they would use or if they would use different categories.

2.4. Item Discrimination

A common threat to validity comes from lack of discrimination. For example, if items, written to represent the model in Table 2, simply ask what priority a teacher gives for each objective, the researcher could easily find that teachers give a high priority to all objectives, given that the official curriculum mandates all objectives. However, it is unreasonable to think that, even with a mandated curriculum, teachers would give every objective the same priority; thus, such a survey would fail to provide discrimination and the argument for validity weakened. Haryani et al. (2019) attempted to avoid this problem by using bipolar items that required the respondent to compare objectives. For example, an item asked for “Critical Thinking” to be ranked with respect to “Problem Solving.” By this method, it was not possible for a respondent to give every objective the same priority. Discrimination was improved and thus validity was improved.

Another strategy for improving validity is to use distractor items. Distractor items represent elements that do not fit with the foundational model. If the survey is valid, respondents will reject the distractor items. Once again consulting Haryani et al. (2019), their model (Table 2) has two entries labeled irrelevant content. The survey built on this model contained distractor items that asked respondents to compare legitimate objectives with irrelevant content. The researchers obtain a further line of validation evidence if the respondents reject the distractor items as per the model.

2.5. Item Clarity

The lack of item clarity can potentially harm validity, and thus another line of validation evidence is that items are clearly written. There are resources that provide conventional wisdom on the wording of items. For example, Krosnick (2018, p. 100-101) suggests that items be simple and direct, containing no jargon or ambiguous words, or emotionally charged words. Items should not contain double clauses.[****] He argues that it is better to avoid negations and important to avoid writing questions that lead the respondent in a particular direction. He suggests that it is a good idea for the researcher or researchers to read their items aloud before finalizing them because hearing an item can help one detect a lack of clarity.

Clarity of expression includes clarity of terms and concepts. The terms and concepts used in an item need to be ones with which respondents are reasonably conversant. Such clarity is rarely a problem for simple terms such as age. A response on age is never going to be exact but it is highly probable that what a respondent records for age will be within an error range of +/- 6 months. That error range is not going to be a problem for most education research. Haryani et al. (2019) used terms that came from the Indonesia national curriculum. These terms are more complicated than “age” and a person unfamiliar with the Indonesia national curriculum could easily misinterpret the terms. However, in a centralized education system where objectives are mandated, Haryani et al. (2019) reasonably assumed that Indonesian teachers in that system are conversant with terms found in that curriculum. On the other hand, researchers can quickly run into trouble if they use terms open to interpretation amongst potential respondents (Smyth, 2016). For example, a Likert item asking how often a teacher uses an inquiry approach to instruction will be subject to a wide range of teacher interpretations. A better approach would be to describe the teaching approach and then ask a how often the respondent might use this approach or one similar to it (see for example Cobern et al., 2014).[††††]

Moreover, even apparently simple words can be potentially troublesome. Redline (2013), for example, found that a survey asking about “shoes” was open to various interpretations. Does the word shoes include boots? Or sandals? In a test of wording, Redline found that an item specifying the meaning of “shoes” returned different responses from an item that didn’t. The better approach is to break the question down into a set of specific questions, such as how many shoes do you have, how many boots do you have, have any sandals do you have, etc. The point is that when writing items, the researcher needs to explore ways of making sure that critical terms in an item will be understood as intended. Even small wording changes can change how respondents interpret and respond to an item. Cobern, Adams, Pleasants, Bentley and Kagumba (2019), for example, got substantially different survey results in a nature of science study when

[****] Often referred to as ‘double-barreled’ items.

[††††] If the researcher decides to use specific examples, such as specific examples of teaching approaches, then those examples need to be based on the theoretical model for the study. For example, see Haryani et al. (2019).

the wording of one item was changed. They also found that a change of item wording can have effects on written responses.

If researchers decide to create their own research instrument it's typically because nothing published meets their particular needs. Nevertheless, researchers are likely to find published surveys that are similar to what they need and it is wise to learn from such published efforts. For example, there are many science attitude surveys. Although none of these may meet a researcher's need, the researcher can still learn from published science attitude items. Moreover, while a perfectly applicable instrument may not be available, it is likely that there are existing questions specific to the interests of the researcher. This is particularly true for questions about demographic information. It is common practice for surveys to include questions about demographics, and example questions are easily available online (e.g., Bhat, 2019; Fryrear, 2016; Rosenberg, 2017). Because the effective wording of survey items is so critical to validity it only makes sense for researchers to learn from published research when writing new items, and to use existing items of known validity when possible.

2.6. Ordering of Items

Once the researcher has finalized a set of items, these items have to be ordered for effective presentation (Smyth, 2016). A researcher may be tempted to give little thought to the order in which items are presented in a survey but that would be a mistake. For example, unless there is a specific reason to group similar items together, grouping similar items runs the risk that the first items in the group will influence the responses to the later items in the group. Unless that is what the researcher wants, similar items need to be dispersed throughout a survey typically by randomly assigning position. The rule of thumb on survey length is that respondent attention wanes toward the end of a long survey. Therefore, any items considered critical are best placed towards the start of a long survey. Demographic questions are often listed at the end of a survey because in academia these items are typically less important than the content items, or at least require less deliberation by the respondent. Unless the research question is *how* demographics relate to the content answers (in which case you need both), better to lose demographic data than to lose data having to do with the main focus of the survey. Another possible criterion for ordering items is the amount of reflection a content item requires. Researchers may wish to place items requiring less reflection earlier in the survey so as to help ease the respondent into the survey. The point is that the importance of ordering items should not be overlooked; it is something the researcher should attend to before finalizing an instrument.

3. PRETESTING FOR ITEM EFFECTIVENESS

When potential respondents read an item, they need to understand the item as per the intention of the researcher. Item effectiveness is a matter of item validity. If a potential respondent does not understand the item as intended by the researcher then the respondent won't actually be responding to what the researcher intended to ask. The lines of evidence for item validity include what has been discussed above: model-based items, appropriate item format, item discrimination, item clarity, and item order. Nevertheless, items should always be pretested (Willis, 2016). Once researchers have finished the ordering of items, the items can be pretested as a whole instrument.

Pretesting begins with an external review of the items. The researcher should always have items read by persons who are similar to those for whom the survey is being constructed (the target population) or who are familiar with the target population. In addition, the researcher needs to have expert readers who are knowledgeable about the subject matter and can read items with respect to content (Dillman, Smyth, and Christian, 2014, p 249-250). The researcher needs to have a set of questions for the external reviewers to think about. For example, external reviewers might be asked:

- Having read the items, what do you think this survey is about?
- Do you think that subjects in our target population [stipulate that population] will have difficulties understanding any of these questions?
- Are there items that you suspect most respondents will answer the same way? In other words, are there items that you suspect will not return a range of responses? That is, most everyone will respond similarly.
- Are there any changes to items that you would recommend making? Changes that would make items more easily understood.
- Is any of the content wrong in your opinion?
- How much time do you think it will take a person in our target population [stipulate that population] to thoughtfully complete the survey?

A researcher might give external reviewers a copy of a survey along with these questions asking the reviewers to respond to the survey items and then to these questions. The researcher uses subsequent feedback for making adjustments to individual items and perhaps the survey as a whole. Or, this feedback could be obtained through interviews (see next section on cognitive interviewing) or focus groups. As noted earlier, the researcher may need to have two types of external reviewers: external reviewers who represent the target population and content expert external reviewers.

Pilot studies can also be useful for evaluating item effectiveness, though typically you would not conduct a pilot study prior to having an instrument externally reviewed. A pilot study involves having a sample from the target population take the survey. The researcher can check the pilot study data for the presence of seriously skewed item responses. Such items fail the objective of having items that discriminate amongst the respondents. Lack of correlation between items represents one kind of problem – weakening the targeted construct; strongly correlated items can mean another kind of problem, indicating too little difference between the items – again, little or no discrimination. If a survey contains filter or distractor items, these can be checked through a pilot study. If these items function as expected, the argument for validity is strengthened.

3.1. Pretesting via Cognitive Interviewing

As noted above, pretesting can also include "cognitive interviews,"

> …an applied approach to identifying problems in survey questionnaires and related materials, with the goal of reducing the associated response errors. … The cognitive interview is conducted using verbal probing techniques, as well as "think-aloud," to elicit thinking about each question (Willis, 2018, p. 103).

Cognitive interviews are critical for surveys that are to include theoretical constructs because item validity rests on respondents understanding the construct intended by the researcher. The goal of interviewing is to determine the likely ways in which respondents from a target population will interpret constructs important to the research. Earlier we gave the example of how respondents can misunderstand the word "shoes." Here is another example. If you ask a person if they own a car, how will they interpret "car"? Does car include small trucks and SUVs? Interviews with persons from the target population would give the researcher at least some insight into how broadly or how narrowly the concept of "car" is likely to be interpreted (Blair & Conrad, 2011).

If concepts such as "cars" and "shoes" are open to various interpretations, just think about the many ways that students or teachers might define the concepts of "teacher centeredness" and "student centeredness," or in science specifically, the concept of "inquiry instruction." A survey could ask science teachers, using a Likert scale, to what extent they think inquiry instruction is effective; but the problem is that you couldn't be sure exactly what the teachers meant by inquiry.

Any time a researcher is considering the use of survey items that include potentially ambiguous concepts, cognitive interviews are critical. It is during an interview that the researcher can learn to what extent a "potentially" ambiguous concept is actually ambiguous. If it turns out that the concept is not ambiguous during an interview, then the researcher can go ahead. However, if the concepts are found to be ambiguous then different strategies are needed for item structure. An interview might yield a narrow range of meanings and in that case an item might indicate this range by, for example, putting a few clarifying terms in parentheses after the concept. Or, instead of writing a single item, the researcher could consider writing a set of items using the various terms uncovered during the interviews. However, the researcher could find it difficult to interpret the results from a set of items ranging around the researcher's intended concept.

Another possibility for dealing with potentially ambiguous concepts is to write scenarios or vignettes that (for the researcher's purposes) represent an intended concept. The idea is that a description of an event communicates more clearly than does a label for an event. For example, questions about a short description (or vignette) of an inquiry lesson (as per the researcher's definition of inquiry) should return more valid responses than merely asking a respondent about inquiry lessons where the definition of "inquiry" is left up to the respondent. Bear in mind that if vignettes or examples are to be used, these also need to be based on the foundational model for the research, otherwise validity is threatened.

Cognitive interviews are not without problems. Government survey labs in the USA make widespread use of cognitive interviews for evaluating public opinion surveys; however, Willis (2018, p. 104) notes that "it is unclear whether or not independent researchers testing the same questionnaire would reach the same conclusions." Willis (2018, p. 104) further notes that little is known about "under what conditions are cognitive interviewing results stable and reliable, and what can researchers do to enhance those conditions." Furthermore, all pretesting is influenced by sample size.

> From the perspective of sample size, a problem's prevalence affects the number of pretest interviews needed to identify it. For example, if we conduct a specified number of cognitive interviews (n) and a particular problem (f) occurs with prevalence (π), what is the probability (P_f) that it will be observed at least once by the nth pretest interview, i.e., at some point in a sample of size n? The probability of observing a problem in the pretest sample depends on two factors: how often the problem occurs (π) and how likely it is to be detected when it does occur (d) (Blair & Conrad, 2011, p 640-641).

Blair and Conrad (2011, p 636) found that,

> Multiple outcome measures showed a strong positive relationship between sample size and problem detection; serious problems that were not detected in small samples were consistently observed in larger samples.

Hence, the difficult question is how many interviews to conduct, because the more interviews one conducts, the less likely it is that the researcher will miss problems. Fortunately for most education researchers, the saturation rule can be used as a guide (Cobern & Adams, 2020; Seidman, 2006). This rule advises interviewing people until the researcher ceases to hear anything new. Admittedly, this role does not guarantee that the researcher won't miss rare opinions but the researcher accepts this risk on the basis that rare misunderstandings of an item will not have a significant impact on the research. Moreover, finding a rare misunderstanding of an item does not necessarily suggest a corrective action. Consider the possibility that a researcher interviews as few as 10 people finding that nine of the 10 understand the item as written. Does the researcher change the item because of the one person out of 10 who misunderstood the question? Probably not. Chances are if the researcher changes the item in light of that one person the other nine might then have difficulties (see Dillman et al., 2014, p. 248). The point is the researcher only needs to interview enough people to be assured that the item is generally understood. Changing an item requires that there be a general misunderstanding or perhaps a significant misunderstanding of an item on the part of several people – not just one.

3.2. In Summary

Surveys should always be pretested and if survey items include potentially ambiguous concepts then the researcher should use cognitive interviews to evaluate such concepts for ambiguity. Whether or not to rewrite an item based on the findings of cognitive interviews is a matter of researcher judgment. Typically, the researcher would not rewrite an item unless the interview findings indicated substantial potential for misinterpretation. If an item is to be rewritten, there are various approaches other than simply changing the words of the item. The researcher can consider using, for example, a set of questions rather than one, adding descriptions to clarify the potentially ambiguous concept, or employing illustrative vignettes in the place of terms.

4. PILOT TESTING FOR RELIABILITY

Having done all of the above in order to have a strong argument for survey validity, there remains the question of how reasonable it is that respondents' responses are stable. Put another way, if you ask respondents the same set of questions two weeks later, will they respond the same way? This stability is what reliability is about (AERA, APA, NCME, 2014, p. 33-47). Survey items are reliable to the extent that responses are stable. The responses don't change over short periods of time during which it is reasonable to assume that nothing has occurred to change respondent views.

Many researchers report Cronbach's alpha as a measure of instrument reliability. Following Taber (2018), we believe this to be a mistake. Cronbach's alpha indicates the internal consistency amongst the group of items. If you have a category, such as "Research Experience" referred to earlier, represented by a set of items, those items need to be highly correlated if they are to validly represent this category. The correlational strength can be gauged using Cronbach's alpha. However, internal consistency is not the same thing as stability over time, which is what reliability is. Hence, a better way to gauge reliability is to give same group of people the instrument twice and then calculate the correlation between two sets of responses (Multon, 2010). The standard benchmark for reliability is that the two episodes of taking the instrument correlate at 0.70 or better. The researcher must bear in mind that testing for reliability is sensitive to the size of the sample. The reliability test-retest will not be effective if the sample is too small. There is no hard and fast rule about how much time should separate the test and retest but conventional wisdom suggests a separation of 10 days to two weeks. There needs to be enough separation so that the first test has faded in the respondent's mind; but the separation cannot be too long because of the risk of intervening factors that would change respondent opinions recorded by the retest.

Finally, the data from a reliability test-retest should also be examined for validity. For example, factor internal consistency should be rechecked, response distributions for items should be rechecked, and the effectiveness of filter or distractor items should be checked.

5. CONCLUSION

As noted at the beginning, this document is a practical guide. There is far more to know about surveys and survey construction than what has been discussed here; this guide should only be used as a starting point. At the very least, researchers using this guide should also consult one or more of the excellent handbooks available on survey research. Finally, researchers should keep research notes about the procedures used for establishing validity and reliability. Such notes are important for informing the argument that a researcher will need when writing for research publication.

Acknowledgements

Not applicable.

Declaration of Conflicting Interests and Ethics

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

ORCID

William W. Cobern https://orcid.org/0000-0002-0219-203X
Betty AJ Adams http://orcid.org/0000-0002-8554-8002

6. REFERENCES

American Education Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cobern, W. W. (2000). *Everyday thoughts about nature: An interpretive study of 16 ninth graders' conceptualizations of nature*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Cobern, W. W., & Adams, B. A. (2020). When interviewing: how many is enough? *International Journal of Assessment Tools in Education, 7*(1), 73-79. https://doi.org/10.21449/ijate.693217

Cobern, W. W., Adams, B. A. J., Pleasants, B. A.-S., Bentley, A., & Kagumba, R. E. (2019, March 31-April 3, 2019). Investigating the potential for unanticipated consequences of teaching the tentative nature of science. Paper presented at the National Association for Research in Science Teaching, Baltimore, MD.

Cobern, W. W., Gibson, A. T., & Underwood, S. A. (1999). Conceptualizations of nature: An interpretive study of 16 ninth graders' everyday thinking. *Journal of Research in Science Teaching, 36*(5), 541-564.

Cobern, W. W., Schuster, D. G., Adams, B., Skjold, B., Mugaloglu, E. Z., Bentz, A., & Sparks, K. (2014). Pedagogy of Science Teaching Tests: Formative Assessments of Science Teaching Orientations. *International Journal of Science Education, 36*(13), 2265-2288. Retrieved from http://bit.ly/RE95xZ

Baker, F. B. (2001). *The basics of item response theory*: ERIC Clearinghouse on Assessment and Evaluation. Retrieved from http://echo.edres.org:8080/irt/baker/final.pdf

Bhat, A. (2019). Top 7 Demographic survey questions for questionnaire. Retrieved from https://www.questionpro.com/blog/demographic-survey-questions/

Blair, J., & Conrad, F. G. (2011). Sample size for cognitive interview pretesting. *Public Opinion Quarterly, 75*(4), 636-658. https://doi.org/10.1093/poq/nfr035

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best Practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*, 149-149. https://doi.org/10.3389/fpubh.2018.00149

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sciences Education, 15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Bruck, A. D., & Towns, M. (2013). Development, implementation, and analysis of a national survey of faculty goals for undergraduate chemistry laboratory. *Journal of Chemical Education, 90*(6), 685-693. https://doi.org/10.1021/ed300371n

Bruck, L. B., Towns, M., & Bretz, S. L. (2010). Faculty perspectives of undergraduate chemistry laboratory: goals and obstacles to success. *Journal of Chemical Education, 87*(12), 1416-1424. https://doi.org/10.1021/ed900002d

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method* (4th Ed). Hoboken, NJ: John Wiley & Sons, Inc.

Fryrear, A. (2016). How to write better demographic survey questions. Retrieved from https://www.surveygizmo.com/resources/blog/how-to-write-better-demographic-questions/

Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: reflections from over 20 years of research. *Educational Psychologist, 50*(1), 14-30. https://doi.org/10.1080/00461520.2014.989230

Haryani, E., Cobern, W. W., & Pleasants, B. A.-S. (2019). Indonesia Vocational High School Science Teachers' Priority Regarding 21st Century Learning Skills in Their Science Classrooms. *Journal of Research in Science Mathematics and Technology Education, 2*(2), 105-133.

Krosnick, J. A. (2018). Improving question design to maximize reliability and validity. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research* (pp. 95-101). New Ydork: Palgrave Macmillan.

Lamb, R. L., Annetta, L., Meldrum, J., & Vallett, D. (2012). Measuring science interest: RASCH validation of the science interest survey. *International Journal of Science and Mathematics Education, 10*(3), 643-668. https://doi.org/10.1007/s10763-011-9314-z

Luo, T., Wang, J., Liu, X., & Zhou, J. (2019). Development and application of a scale to measure students' STEM continuing motivation. *International Journal of Science Education, 41*(14), 1885-1904. https://doi.org/10.1080/09500693.2019.1647472

Multon, K. D. (2010). Test-retest reliability. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1495-1498). Thousand Oaks, California: SAGE Publications, Inc.

Nyutu, E. N., Cobern, W. W., & Pleasants, B. A.-S. (2020). Development of an instrument to assess students' perceptions of their undergraduate laboratory environment. *The Journal for Research and Practice in College Teaching, 5*(1), 1-18. Retrieved from https://journals.uc.edu/index.php/jrpct/article/view/1492

Redline, C. (2013). Clarifying categorical concepts in a web survey. *Public Opinion Quarterly, 77*(S1), 89–105. https://doi.org/10.1093/poq/nfs067

Rosenberg, S. (2017). Respectful collection of demographic data. Retrieved from https://medium.com/@anna.sarai.rosenberg/respectful-collection-of-demographic-data-56de9fcb80e2

Ruel, E. E., Wagner III, W. E., & Gillespie, B. J. (2016). *The practice of survey research: theory and applications*. Los Angeles, CA: SAGE.

Seidman, I. E. (2006). *Interviewing as qualitative research: a guide for researchers in education and the social sciences, 3rd Edition*. New York: Teachers College, Columbia University.

Smyth, J. D. (2016). Chapter 16: Designing Questions and Questionnaires. In C. Wolf, D. Joye, T. W. Smith, & Y.-c. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 218-235). https://doi.org/10.4135/9781473957893

Staus, N. L., Lesseig, K., Lamb, R., Falk, J., & Dierking, L. (2019). Validation of a measure of STEM interest for adolescents. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-019-09970-7

Taber, K. S. (2018). The use of Cronbach's Alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273-1296. Retrieved from https://doi.org/10.1007/s11165-016-9602-2

Vannette, D. L., & Krosnick, J. A. (2018). *The Palgrave handbook of survey research*. New York: Palgrave Macmillan.

Willis, G. B. (2016). Chapter 24: Questionnaire pretesting. In C. Wolf, D. Joye, T. W. Smith, & Y.-c. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 359-380). https://doi.org/10.4135/9781473957893

Willis, G. B. (2018). Cognitive interviewing in survey design: State of the science and future directions. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research* (pp. 103-107). New York: Palgrave Macmillan.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 420–435

https://doi.org/10.21449/ijate.707272

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

The Relationship between Effective Teacher Characteristics and Reasons for Choosing Teaching Profession: Development of an Effective Teacher Inventory

Cetin Toraman [1,*], **Melek Cakmak** [2]

[1] Çanakkale Onsekiz Mart University, Medical School, Medical Education, Çanakkale, Turkey
[2] Gazi University, Faculty of Education, Curriculum & Instruction, Ankara, Turkey

ARTICLE HISTORY

Received: Mar. 21, 2020

Revised: June 29, 2020

Accepted: Aug. 21, 2020

KEYWORDS

Effective teacher characteristics,

Choosing teaching profession as a career,

Inventory development

Abstract: The main purpose of this study is to investigate the relationships between the opinions of secondary school teachers about effective teacher characteristics and their reasons for choosing the teaching profession. In this context, the study first intends to develop a measurement tool to identify effective teacher characteristics. The study is of a correlational research type. Data were collected from three different groups of secondary school teachers. The effective teacher characteristics inventory and the choosing teaching profession as a career scale were used to collect data. The data were analysed using the exploratory and confirmatory factor analyses, Cronbach Alpha internal consistency analysis and multiple linear regression. Given the limitations of the study, groups from which data were obtained, the results of the analyses have shown that the "Effective Teacher Characteristics Inventory" is able to make valid and reliable measurements for effective teacher characteristics under four independent scales (e.g., subject matter knowledge, teaching skills, personality characteristics and professional development). The multiple linear regression has demonstrated that the predictor variables in the model, subject matter knowledge, personality characteristics, and professional development are positive predictors for teaching skills. However, reasons for choosing the profession is not a significant predictor for teaching skills of teachers.

1. INTRODUCTION

Teachers have a significant role and responsibility in the success of an education system. Many studies in the literature have dealt with teachers. Teacher qualifications, effective teacher characteristics and teacher influence comprise a considerable portion of these. The present study focuses on effective teacher characteristics and reasons for choosing teaching as a profession (Griffin, McGaw, & Care, 2012; Metzlera, & Woessmann, 2012).

1.1. Effective Teacher

Although a general review of the literature involving studies on effective teaching would reveal presence of many studies taking the teacher as their theme, they show various differences with respect to their contexts, focal points, methods, and results. Being an important dimension of

CONTACT: Çetin Toraman ✉ toramanacademic@gmail.com Çanakkale Onsekiz Mart Üniversitesi, Medical School, Department of Medical Education, Çanakkale/Turkey

ISSN-e: 2148-7456 /© IJATE 2020

effective teaching as well as the subject of this study, effective teacher characteristics are explained and discussed in the relevant body of literature under various subtopics such as *competency in subject matter knowledge, teaching skills, personality characteristics, and professional development*. All these characteristics have been the focus of various studies (Brophy, 2000; Cotton, 2000; Danielson, 2007; Gholam & Kobeissi, 2012; Goe, Bell & Little, 2008; Jones, Jenkin & Lord, 2006; Kyriakides, Campbell, & Christofidou, 2002; McArdle & Coutts, 2003; McEwan, 2002; Muijs & Reynolds, 2000; Polk, 2006; Saunders, 2000; Shindley Elliott, 2010; Stronge, 2007; Swainston, 2008; Tucker & Stronge, 2005; Woolfolk, 1998).

For subject matter knowledge, which has an important place among effective teacher characteristics, various terms have been used in the literature including good command of the subject matter concepts (Polk, 2006), knowledge of pedagogy (Polk, 2006; Tucker & Stronge, 2005), and knowledge of contents (Shulman, 1986). Woolfolk (1998) has pointed out the importance of the role of knowledge and clarity of the teacher's instructions, explanations, and presentations in students' learning. The relevant body of literature emphasizes a number of characteristics such as establishing a positive classroom setting, effective use of various teaching methods or techniques, presenting the subject by linking it to daily living (Tucker & Stronge, 2005), and coming prepared to classroom (McArdle & Coutts, 2003).

Some personal characteristics of an effective teacher are listed as follows (Kyriakides, Campbell, & Christofidou, 2002; Muijs & Reynolds, 2000; Stronge, 2007; Swainston, 2008; Tucker & Stronge, 2005): Geniality, consistence, self-confidence, honesty, appreciative of student views, ability to communicate effectively, positive attitude, having great expectations from students, accepting student feelings without judgement, setting an example for students, self-reliance, a flexible, creative and tolerant disposition, and a democratic attitude. There are also studies stressing the professional development of an effective teacher (Goe, Bell & Little, 2008; McEwan, 2002; Polk, 2006; Stronge, 2007). According to these studies, effective teachers believe in life-long learning, follow research studies in their profession, appreciate personal development, invest in their own education, and closely monitor opportunities in personal development such as in-service trainings, congresses, and conferences.

1.2. Reasons for Choosing Teaching Profession as a Career

Effective teaching is not limited to having necessary knowledge and skills; it also requires a positive attitude towards the profession and motivation (Heinz, 2015; Watt, Richardson & Wilkins, 2014). At this point, studies become important that deal with what motivates individuals to become a teacher, how they perceive the profession of teaching and what their expectations are from a career development. Studies have explained the reason for choosing teaching as a profession under three categories: a) Extrinsic reasons such as salary and long leaves, b) intrinsic reasons such as interest, personal experience and intellectual satisfaction, and c) altruism such as a desire to contribute to the development of other people (Brookhart & Freeman, 1992; Kyriacou & Coulthard 2000; Moran, Kilpatrick, Abbott, Dallat, & McClune, 2001). Yu (2011) has come up with a more comprehensive list of the factors affecting career choices of teachers including intrinsic, altruistic, and extrinsic reasons, perceived teaching skills, social effect, and teaching experiences.

The results shown by research studies on effective teacher characteristics have an important role in many respects such as teacher education, professional development and assessment of teachers (Stronge, Ward & Grant, 2011). In defining the knowledge and skills needed by teacher candidates, ensuring professional development, making valid and reliable assessments of teachers (Stronge, Ward, & Grant, 2011, p.339 as cited in Darling-Hammond & Bransford, 2005; Hanushek, 2008; National Academy of Education, 2008) and in many other context that can be listed, identification of effective teacher characteristics on the basis of teaching levels is important. Although there are many studies carried out at various levels in this subject in the relevant literature (Gholam & Kobeissi, 2012; Keeley, Smith, & Buskist, 2006; Moran, 2005; Shindley Elliott, 2010), these studies have been conducted mostly with teacher candidates with limited number of studies dealing with this subject at primary and secondary school levels.

In the literature review, there is no study investigating the relationships between the effective teacher characteristics and the reasons for choosing teaching. It can be said that this situation inspired the research. The main purpose of this study is to investigate the relationships between the opinions of secondary school teachers on effective teacher characteristics and the reasons why they choose the teaching profession. In this context, the study first intends to develop a measurement tool to identify effective teacher characteristics.

2. METHOD

This study is structured as a descriptive research because it describes the features of the measuring tool under development and as a correlational research (Fraenkel, Wallen, & Hyun, 2012) in the sense that it questions the relationships between effective teacher characteristics and reasons for choosing the teaching profession.

2.1. Participants

The study data were obtained from three different groups. The first group consisted of teachers working at secondary school level (n=421). Data were collected from this participating group for the purpose of obtaining information about the construct validity and reliability level of the *Effective Teacher Characteristics Inventory* that was planned to serve as a measurement tool in this study. The second group was again formed of teachers working at secondary schools (n=403). Data were collected from this second group to test whether or not the construct of the *Effective Teacher Characteristics Inventory* as a measurement tool developed for this study is verified. The last group from which data were collected in the study consisted of secondary school teachers (n=321) and the data were collected from this group for the purpose of exploring the relationships between effective teacher characteristics and reasons for choosing the teaching profession. These three different groups were formed using the purposive sampling method, a sampling method for unknown probabilities. In non-probability sampling methods, the probability of selecting each person from the population to the sample cannot be calculated (Sumbuloglu & Sumbuloglu, 2005). Convenience sampling is based on working with a portion of the population, not the whole (Senol, 2012). When using convenience sampling, researchers determine the characteristics of those who will comprise the study population and try to reach the persons who have these characteristics. Some variables of the participants are shown in Table 1 and 2.

Table 1. *Distribution of teachers in exploratory factor analysis and confirmatory factor analysis groups according to various variables*

Group of exploratory factor analysis				Group of confirmatory factor analysis			
	Variable	f	%		Variable	f	%
Gender	Female	292	69,4	Gender	Female	290	72
	Male	129	30,6		Male	113	28
	Total	421	100		Total	403	100
Experience	1-5 years	10	2,4	Experience	1-5 years	8	2
	6-10 years	45	10,7		6-10 years	43	10,7
	11-15 years	124	29,5		11-15 years	110	27,3
	16-20 years	147	34,9		16-20 years	142	35,2
	21 years and over	95	22,6		21 years and over	100	24,8
	Total	421	100		Total	403	100
Graduated Faculty	Faculty of Education	315	74,8	Graduated Faculty	Faculty of Education	303	75,2
	Other	106	25,2		Other	100	24,8
	Total	421	100		Total	403	100

Subject Matter /Areas of Expertise	Turkish	90	21,4	Subject Matter /Areas of Expertise	Turkish	83	20,6
	Mathematics	56	13,3		Mathematics	49	12,2
	Science Education	56	13,3		Science Education	59	14,6
	Social Sciences	40	9,5		Social Sciences	37	9,2
	English	40	9,5		English	37	9,2
	Psychological counseling and guidance	24	5,7		Psychological counseling and guidance	22	5,5
	Music	16	3,8		Music	18	4,5
	Visual arts	25	5,9		Visual arts	25	6,2
	Physical Education	13	3,1		Physical Education	13	3,2
	Technology Design	26	6,2		Technology Design	25	6,2
	Informatics/Information Technology	14	3,3		Informatics/Information Technology	14	3,5
	Theology	21	5		Theology	21	5,2
	Total	421	100		Total	403	100

The majority of teachers involved in scale development groups are women (69-72%). About half of the teachers have an experience of 11-20 years. Teachers from different subject matter/areas of expertise at secondary school level are included in this group.

Table 2. *Distribution of teachers in relational modeling groups according to various variables*

	Variable	f	%
Districts of Ankara where she/he works	Çankaya	54	16,8
	Mamak	51	15,9
	Yenimahalle	62	19,3
	Keçiören	51	15,9
	Altındağ	57	17,8
	Sincan	46	14,3
Gender	Female	220	68,5
	Male	101	31,5
Experience	1-5 years	15	4,7
	6-10 years	49	15,3
	11-15 years	93	29
	16-20 years	105	32,7
	21 years and over	59	18,4
Subject Matter /Areas of Expertise	Turkish	55	17,1
	Mathematics	48	15
	Science Education	45	14
	Social Sciences	36	11,2
	English	25	7,8
	Psychological counseling and guidance	22	6,9
	Music	17	5,3
	Visual arts	16	5
	Physical Education	21	6,5
	Technology Design	12	3,7
	Informatics/Information Technology	10	3,1
	Theology	14	4,4
	Total	321	100

2.2. Data Collection Instruments

Two different data collection tools were used in this study. The first of these data collection tools, the "*Effective Teacher Characteristics Inventory*" has been developed by the researchers. A pool of items was constructed as the first step in developing the inventory. When creating this pool, information obtained from the literature and information obtained as a result of the Delphi process that was conducted by the investigators were used. During the Delphi study, the question "What are effective teacher characteristics?" was asked to 139 teachers working at secondary schools, 402 secondary school students and 204 students from faculties of education. Additionally, opinions of 14 teacher educators working at various universities were obtained. The effective teacher characteristics stated by all participants were listed as a result of a content analysis. The characteristics listed were first sent to a group of four experts from the field of Curriculum & Instruction and one from the field of Guidance & Psychological Counselling and their views on the characteristics were obtained. The final version of the effective teacher characteristics that were corrected and redesigned based on the views received were sent again to the same experts by mail. After taking the latest suggestions into consideration, the item pool for effective teacher characteristics was finalized and then administered. The groups that collected data in the Delphi process, the groups where data was collected to scale development process, and the group where data was collected for relationship analysis were formed from different participants.

As a result of Delphi process and expert opinions, a list of effective teacher features consisting of 80 items was reached. 80 items formed the item pool to develop the scale. These items are structured in likert type before being implemented. As explained in detail in the results/findings section, an inventory of 25 items and four independent scales was obtained from the 80 items pool.

"*Effective Teacher Characteristics Inventory*" is able to make valid and reliable measurements for effective teacher characteristics under four independent scales (subject matter knowledge, personality characteristics, professional development, and teaching skills). Subject matter knowledge is a scale of four items. The lowest score that can be obtained from this scale is 4, the highest score is 20. Personality characteristics is a scale of seven items. The lowest score that can be obtained from this scale is 7, the highest score is 35. Professional development is a scale of 4 items. The lowest score that can be obtained from this scale is 4, the highest score is 20. Teaching skills consist of three sub-scales and 10 items. The lowest score that can be obtained from this scale is 10, the highest score is 50.

The other measurement tool used in the study was the "Choosing Teaching Profession as a Career Scale". The Choosing Teaching Profession as a Career Scale was developed by Lai, Chan, Ko, & So (2006) and adapted to Turkish by Balyer & Ozcan (2014). The Turkish version of the scale shows that the scale consists of 20 items and 3 subdimensions. These 3 subscales are: "Altruistic/intrinsic reasons, extrinsic reasons and influence of others". Balyer & Ozcan (2014) conducted their study with a total of 1410 faculty of education students from 8 different state universities and 220 students took part in performing the validity and reliability analyses. The CFA results of the Turkish version of the scale were; X^2/sd=2,3, GFI=0,90, AGFI=0,80, NFI=0,95, NNFI=0,95, CFI=0,92, RMR=0,10, RMSEA=0,08, and SRMR=0,09, which were at an acceptable level according to the literature. The Cronbach alpha coefficients of the scale were 0.91 for the altruistic/intrinsic reasons subdimension, 0.80 for the extrinsic reasons subdimension and 0.74 for the influence of others subdimension. Since the scale, which had been adapted by Balyer & Ozcan for teacher candidates, was meant to be used for secondary school teachers in this study and its target population changed, it was separately tested on secondary school teachers (n=321) who would be subject to the last administration in this study to show if it would work with the same structure on teachers. This testing was done with CFA. The fit indices obtained were RMSEA=0,077, RMR=0,022, GFI=0,951, AGFI=0,904, NFI=0,911, IFI=0,918, CFI=0,956, and X^2/sd=2.87, which were within the limits of acceptable values. Cronbach Alpha value is the basis for the reliability of the scale as internal consistency. When the Cronbach alpha reliability coefficients of the scale were calculated, the reliability coefficients for the teacher version were found to be 0.89 for the altruistic/intrinsic reasons subdimension, 0.77 for the extrinsic reasons subdimension and 0.76 for the influence of others subdimension.

2.3. Data Analysis

Missing values were not found in the data file. Therefore, it was decided to apply factor. The principles competent method was used in the factor analysis. Whether the data set was suitable for a factor analysis was tested with Kaiser Meyer Olkin (KMO) and Bartlett's Test of Sphericity value. KMO is a criterion relating to the sufficiency of sampling. The KMO statistic ranged between 0 and 1. A KMO value less than 0.500 is usually unacceptable and may necessitate collection of more data. Values between 0.500 and 0.700 are accepted as moderate, between 0.700 and 0.800 as good, between 0.800 and 0.900 as very good and those over 0.900 as excellent (Cokluk, Sekercioglu & Buyukozturk, 2010; Field, 2018; Tabachnick & Fidell, 2013). The Bartlett's Test of Sphericity tests whether the variance-covariance matrix is proportional to a defined matrix. If the test result is significant, it is considered as a global and multivariate normality. However, a disadvantage of this test is that it is influenced by the sample size. With larger samples, the probability of the result to turn out significant increases (Cokluk, Sekercioglu, & Buyukozturk, 2010; Tabachnick & Fidell, 2013). The fit indices in the analysis results obtained for the confirmatory factor analysis (CFA) were reviewed. The results of the fit indices searched in the literature as reference are shown in Table 3.

Table 3. *Fit index reference values accepted for CFA*

Fit-index	Acceptable Limits	Perfect Fit Limits	Source
RMSEA (Root mean Square Error of Approximation)	0.05≤RMSEA≤0.08	0≤RMSEA≤0.05	Hooper, Coughlan, & Mullen, 2008; Hu, & Bentler, 1999; Simsek, 2007; Vieira, 2011
RMR (Root Mean Square Residual)	0.05<RMR≤0.08	0≤RMR≤0.05	Anderson, & Gerbing, 1984; Hooper, Coughlan, & Mullen, 2008; Hu, & Bentler, 1999; Kline, 2005; Marsh, Balla, & McDonald, 1988
GFI (Goodness of Fit Index)		0.90 and over	Hooper, Coughlan, & Mullen, 2008; Kline, 2005
AGFI (Adjusted Goodness of Fit Index		0.90 and over	Anderson, & Gerbing, 1984; Hooper, Coughlan, & Mullen, 2008; Kline, 2005; Marsh, Balla, & McDonald, 1988
NFI (Normed Fit Index)		0.95 and over	Bentler, 1990; Cokluk, Sekercioglu, & Buyukozturk, 2010; Hu, & Bentler, 1999; Kline, 2005; Simsek, 2007
IFI (Incremental Fit Index)	0.90≤IFI≤0.94	0.95 and over	Bentler, 1990; Cokluk, Sekercioglu, & Buyukozturk, 2010; Hu, & Bentler, 1999; Simsek, 2007
CFI (Comparative Fit Index)	0.90≤CFI≤0.94	0.95 and over	Bentler, 1990; Cokluk, Sekercioglu, & Buyukozturk, 2010; Hooper, Coughlan, & Mullen, 2008; Hu, & Bentler, 1999; Simsek, 2007
X^2/sd	$2<X^2$/sd≤5	$0≤X^2$/sd≤2	Kline, 2005; Ozdamar, 2016; Tabachnick, & Fidell, 2013

The regression analysis was planned to be performed with a "*Multiple Linear Regression*" (Ozdamar, 2013). For this reason, the normality of data distribution was tested. A Kolmogorov Smirnov normal distribution test showed that the data were not normally distributed (*p*<.05). Tests testing normality are excessively sensitive (Tabachnick, & Fidell, 2013). In many studies (especially in social sciences), measurements of dependent variables do not show normal distribution (Pallant, 2016). The Central Limit Theorem argues that if the sample is sufficiently large (n=30+), the distribution of means in the sample will be normal regardless of the distribution of variables and a violation of normal distribution will not cause a big problem (Everitt, & Howell, 2005; Field, 2018; Pallant, 2016; Tabachnick, & Fidell, 2013). Therefore, the deviation in large samples does not depart from the normal considerably. Positive kurtosis tends to disappear in a sample size larger than 100 and negative kurtosis in a sample size larger than 200 (Tabachnick, & Fidell, 2013). In the light of this information, the data was assumed to have a normal distribution and a multiple linear regression analysis was used. VIF statistic was investigated in multiple linear regression. The VIF statistic shows a multiple linear dependency/connection between exploration variables. If the VIF value is close to 1, there are no multiple linear dependencies between the predictor variables (Ozdamar, 2013). Also, in this study, there were no multiple linear dependencies at a high level between the predictor variables. Exploratory factor analysis (EFA), reliability analysis and multiple regression analysis were performed with SPSS. CFA was performed with AMOS.

3. RESULTS/FINDINGS

3.1. Process of Developing a Measurement Tool: Effective Teacher Characteristics Inventory

The structure expected to appear from the 80 items in the item pool considered collectively was tested. As a result of the EFA performed using the principle component method, the measurement tool assumed a 21-factor structure. From the dataset analysis values, KMO was found above 0.500 and Bartlett's value significant (*p*<.05). These values are sufficient according to Field (2018), Kalayci (2005) and Ozdamar (2013). In an effort to reduce the number of factors and find a simpler solution, the scree plot of the factor analysis was examined and it was decided to repeat the factor analysis with three distinct factors where the slope was steepest. As a result of the factor analysis performed by limiting the number of factors to three, 27 items were removed from the scale and a 53-item structure was obtained. However, this structure could not be verified by CFA. Therefore, expert views were obtained from a professor and an associate professor from the Department of Educational Assessment and another associate professor from the Department of Curriculum & Instruction. The experts reviewed the results of the factor analysis. They suggested that the measurement tool was more of an inventory type and each dimension should be considered as a separate measurement tool in line with the groupings of effective teacher characteristics in the literature and made subject to a factor analysis individually. The factor analyses carried out in line with these suggestions revealed that the inventory had four different scales independent of each other. A confirmatory factor analysis showed that these scales had covariances with each other and failed to confirm a scale structure. Thus, the scales remained independent. When there are scales independent of each other in a measurement tool, such measurement tool is referred to as an inventory. Aiken (1997, p. 201) has reported that inventories are designed to measure certain variables through the subsets of the items and a score is obtained from the responses given to a certain subset of the items of an inventory. The extraction values obtained for the four independent scales of the inventory from the factor analyses and the item-total correlations obtained from the reliability analysis are shown in Table 4.

Table 4. *Subscale extraction values and item-total correlations of the effective teacher characteristics inventory*

	Items	Extraction Value	Item-Total Correlation
Scale of Competency of Subject Matter Knowledge	I1: When necessary, I give details of the information on the subject in my class.	0.756	0.723
	I2: I respond to student questions requiring additional information (elaboration/detailing).	0.781	0.753
	I3: I direct my students to sources from which they can obtain additional information on the subject.	0.683	0.665
	I5: I utilize diverse examples related to the subject.	0.420	0.466
Scale of Teaching Skills	I8: I use various assessment methods and techniques.	0.465	0.480
	I11: I use appropriate learning strategies (repetition, review, concept maps, etc.).	0.541	0.714
	I12: I use appropriate teaching strategies (via invention, presentation, etc.).	0.720	0.549
	I13: I apply teaching principles (from concrete to abstract, establishing links with life, from near to distant, from easy to difficult, etc.) in my class.	0.644	0.626
	I15: I take into consideration individual differences of students.	0.650	0.560
	I16: I repeat subjects not understood.	0.673	0.560
	I27: I use reinforcers in appropriate variety and frequency.	0.557	0.561
	I31: I motivate my students.	0.630	0.680
	I32: I use classroom management approaches.	0.789	0.543
	I33: I display democratic behaviour in my class.	0.622	0.732
Scale of Personality Characteristics	I46: I treat fairly in class.	0.510	0.599
	I48: I respect my students.	0.629	0.689
	I52: I display positive attitude towards my students.	0.570	0.645
	I54: I am honest to my students.	0.559	0.642
	I58: I am responsible.	0.571	0.649
	I62: I am open to criticism.	0.447	0.555
	I66: I am sincere (openhearted) to my students.	0.511	0.600
Scale of Professional Development	I70: I appreciate professional development.	0.562	0.551
	I72: I follow novelties.	0.665	0.630
	I73: I follow updates.	0.643	0.608
	I76: I have a tendency to life-long learning.	0.526	0.524

A review of Table 2 shows that the factor analysis item extraction value is above 0.40 and the item-total correlation above 0.450 in the items included in the subscales of the inventory. Factor analysis item extraction and item-total correlation values are at the desired level according to the literature (Cokluk, Sekercioglu & Buyukozturk, 2010; Tabachnick & Fidell, 2013). From the four different scales, only the “Scale of Teaching Skills” has three subfactors within itself. These are monitoring and assessment skills, teaching skills and classroom management skills of the teacher.

Four items in the scale of competency of subject matter knowledge were found to explain 66% of the characteristic in question, the scale of teaching skills 63% of the characteristic in question (the remaining items in three-factor structure), the scale of personality characteristics 54% of the characteristic in question, and the scale of professional development 60% of the characteristic in question. The Cronbach Alpha reliability coefficient was found to be 0.82 for the scale of competency of subject matter knowledge, 0.74, 0.72 and 0.74 for the three factors in the scale of teaching skills, 0.86 for the scale of personality characteristics, and 0.77 for the scale of professional development. The CFA results are shown in Figure 1.

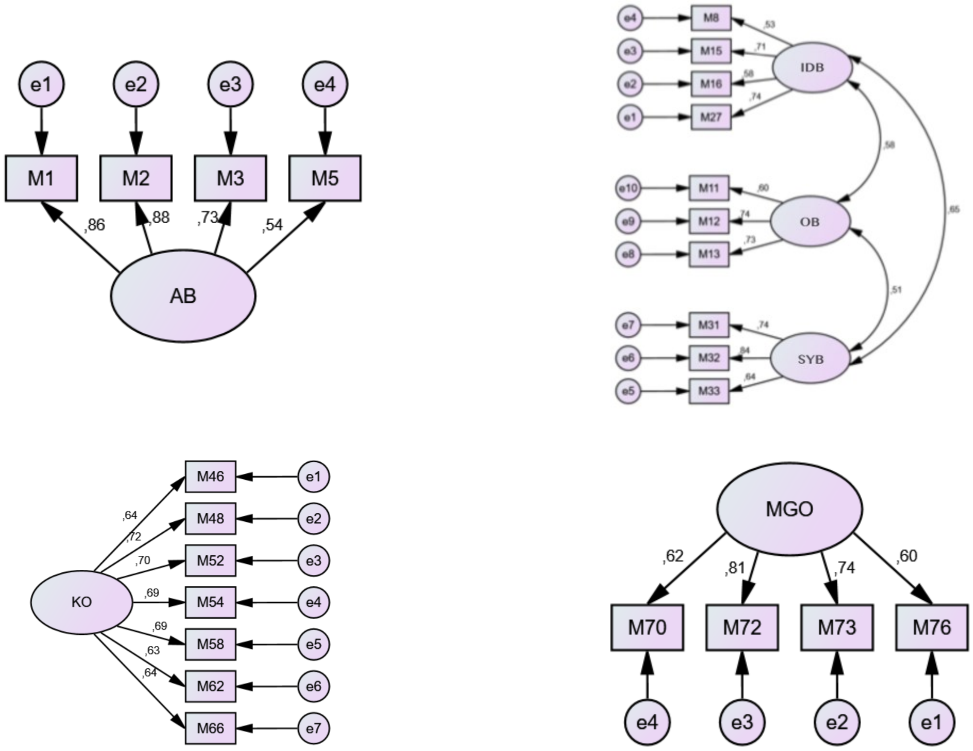


Figure 1. *CFA results of effective teacher characteristics inventory subscales (standardized values) AB: Scale of Competency of Subject Matter Knowledge, IDB: Assessment Skills, OB: Teaching Skills, SYB: Classroom Management Skills, KO: Scale of Personality Characteristics, MGO: Scale of Professional Development*

The fit indices obtained from CFA diagrams are shown in Table 5.

Table 5. *Fit indices*

Scale	RMSEA	RMR	GFI	AGFI	NFI	IFI	CFI	X^2/sd
Scale of Competency of Subject Matter Knowledge	0.078	0.015	0.993	0.963	0.960	0.970	0.969	1.905
Scale of Teaching Skills	0.071	0,030	0,942	0,900	0,896	0.919	0.918	2.794
Scale of Personality Characteristics	0.069	0.032	0.952	0.905	0.931	0.944	0,944	2.786
Scale of Professional Development	0.074	0.013	0.992	0.960	0.986	0.990	0.990	2.146

Table 3 shows that the fit indices are within excellent and acceptable ranges according to the literature on scale development and the reference values given in Table 1. In the light of these results it can be said that within the limitation of the study groups from which the data were obtained the "Effective Teacher Characteristics Inventory" is capable of making valid and reliable measurements for effective teacher characteristics under four independent scales.

3.2. Variables Predicting Teaching Skills of Teachers

Among the basic skills expected of teachers as professionals, teaching skills have an important role. For this reason, the effects of subject matter knowledge, personality characteristics, professional development, and reasons for choosing the teaching profession on teaching skills were dealt with in this section of the study. To this end, a multiple linear regression analysis was performed. The regression formula tested in the analysis is given below.

$$\begin{aligned}\hat{Y}_{Teaching\ Skills} = {} & b_0 + b_{Subject\ Matter\ Knowledge}X_{Subject\ Matter\ Knowledge} \\ & + b_{Personality\ Characteristics}X_{Personality\ Characteristics} \\ & + b_{Professional\ Development}X_{Professional\ Development} \\ & + b_{Altruist\ Reasons\ to\ Choose}X_{Altruist\ Reason\ to\ Choose} \\ & + b_{External\ Reasons\ to\ Choose}X_{External\ Reasons\ to\ Choose} \\ & + b_{Reasons\ for\ Being\ Affected}X_{Reasons\ for\ Being\ Affected}\end{aligned}$$

The above formula was tested with a multiple regression analysis. Each regression is a model. Therefore, in regression analyses, first a summary and fit of the regression model needs to be shown. A summary of the multiple linear regression model used is shown in Table 6.

Table 6. *Model summary*

R	R^2	Adjusted R^2	Standart Error
0.670	0.449	0.438	3.83

The R^2 value in Table 4 gives information about the exploration rate of the model. Assuming that they affect teaching skills in this model, subject matter knowledge, personality characteristics, professional development, and reasons for choosing the teaching profession as a career (altruistic/intrinsic reasons, extrinsic reasons, influence of others) were included in the model as predictor variables. The predictor variables were found to explain 45% of the variance (R^2=0.449) in teaching skills. The fit values of the model are given in Table 7.

Table 7. *Model fit*

Model	Sum of Square	df	Mean Square	F	*p*
Regression	3750.763	6	625.127	42.597	0.000
Residual	4608.047	314	14.675		
Total	8358.810	320			

The result of an ANOVA test on the fit values of the model in Table 5 was found to show model fit ($F_{(6-314)}$=42.597; *p*<.05). After establishing model exploration rate and model fit, the regression coefficients and prediction levels of the predictor variables were studied. The results are shown in Table 8.

Table 8. *Effect of subject matter knowledge, personality characteristics, professional development, and reasons for choosing the teaching profession as a career on teaching skills of teachers*

Model	B	Std. Error	*t*	*p*	VIF
Constant	13.613	2.023	6.730	0.000	
Subject Matter Knowledge	0.552	0.097	5.711	0.000	1.281
Personality Characteristics	0.284	0.061	4.656	0.000	1.484
Professional Development	0.427	0.086	4.946	0.000	1.540
Altruistic/Intrinsic Reasons	0.066	0.044	1.495	0.136	1.625
Extrinsic Reasons	-0.017	0.039	-0.426	0.670	1.614
Influence of Others	0.110	0.061	1.796	0.073	1.500

A review of Table 6 reveals that the constant was significant. This can be interpreted that some variables not included in the model besides the predictor variables (subject matter knowledge, personality characteristics, etc.) that have been included are also predictors of teaching skills of teachers. From the predictor variables in the model, subject matter knowledge, personality characteristics and professional development are positive predictors of teaching skills of teachers (p<.05). As teachers improve their subject matter knowledge, personality characteristics and professional development, their teaching skills also improve. However, reasons for choosing the profession is not a significant predictor of teaching skills of teachers (p>.05).

4. DISCUSSION and CONCLUSION

Given the limitations of the study groups from which data were obtained, the results of the analyses made in the study have shown that the "*Effective Teacher Characteristics Inventory*" is able to make valid and reliable measurements for effective teacher characteristics under four independent subscales (subject matter knowledge, personality characteristics, professional development, and teaching skills). From the predictor variables in the multiple regression analysis model, subject matter knowledge, personality characteristics and professional development are significant positive predictors of teaching skills of teachers (p<.05). As teachers improve their subject matter knowledge, personality characteristics and professional development, their teaching skills also improve. However, reasons for choosing the profession is not a significant predictor of teaching skills of teachers (p>.05). The results of this study have shown that reasons for choosing the profession is not a significant predictor of teaching skills of teachers. Looking at the literature, some similar studies can be seen. For example, Rots, Aelterman, Devos, & Vlerick (2010) have tested their hypothetical teacher education model on a group of students (n=436) and a group of newly graduated teachers (n=251). In their study, the data were collected using the "Teachers' Sense of Efficacy Scale" developed by Tschannen Moran & Woolfolk Hoy (2001), which included content knowledge, subject matter knowledge, efficacy in classroom management and efficacy in student engagement. The results of their study demonstrated that all values measured by the scale were moderately correlated with the other components included in the model and affected the decision whether to actually perform the teaching profession. Their results point out findings that are different from the results of the present study. This may have been influenced by the specific objective, method, context and timing of the study and other reasons. Tschannen Moran & Woolfolk Hoy (2007) conducted another study on the efficacy of teachers including their teaching skills with teachers who were in their first year of the profession and those who were experienced. The study results have shown that teachers need increasingly more support in the process of their experience in the profession to be able to feel more competent in teaching skills.

In a study of Levine (2017), close to a thousand teacher candidates were asked to list the "characteristics they thought mathematics teachers working at primary education level should have". The list prepared from the opinions of teacher candidates revealed that "patience and content knowledge in mathematics" was one of the top items. Levine interpreted this result that teacher candidates had the thought that they should have content knowledge -competence in subject matter knowledge- for effective teaching when they were still students. Supporting this finding, the results of many studies in the literature (Blömeke, Busse, Kaiser, König & Suhl, 2016; Brewer & Goldhaber, 2000; Kamamia, Ngugi & Thinguri, 2014; Monk, 1994; Monk & King, 1994; Rowan, Chiang & Miller, 1997) show that the competency of teachers and teacher candidates in -subject matter knowledge- has a positive effect on their academic achievement.

The results of another study made by Richorson & Watt (2006) with teacher candidates studying in faculties of education of three large state universities in Australia revealed that "beliefs in teaching skills, value of teaching profession with respect to personal and social benefit and previous learning and teaching experiences" were primarily effective in their choice of the teaching profession.

Their result suggesting that competence in subject matter knowledge, personality characteristics and professional development support teachers' teaching skills seems similar to those found in the literature. The finding in the present study that reasons for choosing the teaching profession was not a significant predictor of teaching skills was not compatible with the literature. In this respect, further studies in Turkey may choose to deal with the relationship between reasons for selecting the teaching profession and effective teacher characteristics. The results of this study can be summarized as follows: A measurement tool called "Effective Teacher Characteristics Inventory" was developed for secondary school level during the study. This tool was in the form of an inventory consisting of four scales independent of each other, namely "competency in subject matter knowledge", "teaching skills", "personality characteristics" and "professional development". The total scores obtained from each independent scale cannot be summed up to obtain an overall total score. Nevertheless, given the present structure of the inventory and the data obtained from this study, it can be considered as a valid and reliable measurement tool.

Another result obtained from this study was that improved subject matter knowledge, personality characteristics and professional development of the teachers also improved their teaching skills. However, reasons for choosing the profession had no impact on teaching skills of the teachers.

Further studies on different samples repeating the validity and reliability testing of the inventory and new validity and reliability evidences to be obtained will further strengthen the technical aspects of the inventory. Additionally, the inventory can be experimented at different levels (primary education, secondary education, higher education) and new validity and reliability evidences can be obtained.

The results of this study have shown that from the effective teacher characteristics, subject matter knowledge, teaching skills, personality characteristics, and professional development were associated with themselves. This result can be taken into consideration in teacher education programs and can contribute significantly to teacher candidates in their effort to get prepared for the profession.

Acknowledgements

This article is extracted from doctoral dissertation entitled "A Research on Effective Teacher Qualities of Middle School Teachers and Their Reasons for Choosing Teaching Profession as a Career" (Advisor: Prof. Dr. Melek Çakmak, Gazi University, Institute of Educational Sciences, Ankara/Turkey, 2019)

Declaration of Conflicting Interests and Ethics

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

ORCID

Çetin Toraman https://orcid.org/0000-0001-5319-0731
Melek Çakmak https://orcid.org/0000-0002-3371-4937

5. REFERENCES

Aiken, L. R. (1997). *Questionnaires and inventories, surveying, opinions and assessing personality*. The USA: John Willey & Sons Inc.

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness of fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*(2), 155-173. https://doi.org/10.1007/BF02294170

Balyer, A., & Ozcan, K. (2014). Choosing teaching profession as a career: Students' reasons. *International Education Studies, 7*(5), 104-115. https://doi.org/10.5539/ies.v7n5p104

Bentler P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Blömeke, S., Busse, A., Kaiser, G., König, J., & Suhl, U. (2016). The relationship between content-specific and general knowledge and skills. *Teaching and Teacher Education, 56*, 35-46. https://doi.org/10.1016/j.tate.2016.02.003

Brewer, D., & Goldhaber, D. D. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*, 129-145. https://doi.org/10.3102/01623737022002129

Brookhart, S. M., & Freeman, D. J. (1992). Characteristics of entering teacher candidates. *Review of Educational Research, 62*(1), 37-60. https://doi.org/10.3102/00346543062001037

Brophy, J. (2000). *Teaching*. Educational Practices Series 1. Switzerland: International Bureau of Education (ERIC Database, ED 440 066). Retrieved from: https://files.eric.ed.gov/fulltext/ED440066.pdf

Cokluk, O., Sekercioglu, G., & Buyukozturk, S. (2010). *Sosyal bilimler için çok değişkenli istatistik (Multivariate statistics for social sciences)*. Ankara: Pegem Akademi

Cotton, K. (2000). *The schooling practices that matter most*. Office of Educational Research and Improvement, USA: Washington, DC (ERIC Database, ED 469 234). Retrieved from: https://files.eric.ed.gov/fulltext/ED469234.pdf

Danielson, C. (2007). *Enhancing professional practice, a framework for teaching*. USA: Association for Supervision and Curriculum Development (ASCD)

Everitt, B. S., & Howell, D. C. (2005). *Encyclopedia of statistics in behavioral science*. The UK: John Willey and Sons

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics*. The USA: Sage

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education*. USA: McGraw Hill

Gholam, A. P., & Kobeissi, A. H. (2012). *Teacher evaluation instruments/systems in lebanon and other major arab countries in comparison to evidenced-based characteristics of effective teacher evaluation instruments*. (Doctoral Dissertation). Graduate Faculty of Saint Louis University, Saint Louis.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. National Comprehensive Center for Teacher Quality. USA: Washington

Griffin, P., McGaw, B., & Care, E. (2012). The changing role of education and schools. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 1-16). Dordrecht, Germany: Springer Science+Business Media B.V. http://dx.doi.org/10.1007/978-94-007-2324-5_2

Heinz, M. (2015). Why choose teaching? An internationalreview of empirical studies exploring student teachers' career motivations and levelsof commitment to teaching. *Educational Research and Evaluation, 21*(3), 258-297. https://doi.org/10.1080/13803611.2015.1018278

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods, 6*(1), 53-60. Retrieved from: www.ejbrm.com/issue/download.html?idArticle=183

Hu L. T., & Bentler P. M. (1999). Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Jones, J., Jenkin, M., & Lord, S. (2006). *Developing effective teacher performance*. London: Paul Chapman

Kalayci, S. (2005). *SPSS uygulamalı çok değişkenli istatistik teknikleri (Multivariate statistics techniques with SPSS applied)*. Ankara: Asil

Kamamia, L. N., Ngugi, N. T., & Thinguri, R. W. (2014). To establish the extend to which the subject mastery enhances quality teaching to student-teachers during teaching practice. *International Journal of Education and Research, 2*(7), 641-648. Retrieved from: https://www.ijern.com/journal/July-2014/51.pdf

Keeley, J., Smith, D., & Buskist, W. (2006). The teacher behaviors checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology, 33*(2), 84-91. https://doi.org/10.1207/s15328023top3302_1

Kline, T. J. B. (2005). *Psychological testing, a practical approach to design and evaluation*. The USA: Sage

Kyriacou, C., & Coulthard, M. (2000). Undergraduates' views of teaching as a career choice. *Journal of Education for Teaching, 26*(2), 117-126. https://doi.org/10.1080/02607470050127036

Kyriakides, L., Campbell, R. J., & Christofidou, E. (2002). Generating criteria for measuring teacher effectiveness through a self-evaluation approach: A complementary way of measuring teacher effectiveness. *School Effectiveness and School Improvement, 13*(3), 291-325. https://doi.org/10.1076/sesi.13.3.291.3426

Lai, K. C., Chan, K. W., Ko, K. W., & So, K. S. (2005). Teaching as a career: A perspective from Hong Kong senior secondary students. *Journal of Education for Teaching, 31*(3), 153-168. https://doi.org/10.1080/02607470500168974

Levine, G. (2017). *Effective teacher characteristics: Future teachers' voices*. NERA Conference Proceedings. Retrieved from: http://opencommons.uconn.edu/nera-2017/5

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness of fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*(3), 391-410. https://doi.org/10.1037/0033-2909.103.3.391

McArdle, K., & Coutts, N. (2003) A strong core of qualities-A model of the professional educator that moves beyond reflection. *Studies in Continuing Education, 25*(2), 225-237. https://doi.org/10.1080/0158037032000131547

McEwan, E. K. (2002). *10 traits of highly effective teachers, how to hire, coach, and mentor successful teachers*. The USA: Corwin Press, Inc.

Metzlera, J., & Woessmann, L. (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of Development Economics, 99*(2), 486-496. https://doi.org/10.1016/j.jdeveco.2012.06.002

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13*, 125-145. https://doi.org/10.1016/0272-7757(94)90003-5

Monk, D., & King, J. (1994). Multilevel teacher resource effects on pupil performance in secondary mathematics and science. In Ronald G. Ehrenberg (Ed.), *Choices and consequence* (pp. 29-58). Ithaca, NY: ILR.

Moran, C. (2005). *Teacher and principal perceptions of dispositional characteristics needed by middle school teachers to be most effective in the classroom*. (Doctoral Dissertation). Indiana State University, Indiana

Moran, A., Kilpatrick, R., Abbott, L., Dallat, J., & McClune, B. (2001). Training to teach: Motivating factors and implications for recruitment. *Evaluation & Research in Education, 15*(1), 17-32. https://doi.org/10.1080/09500790108666980

Muijs, D., & Reynolds, D. (2000). School effectiveness and teacher effectiveness in mathematics: some preliminary findings from the evaluation of the mathematics enhancement programme (primary). *School Effectiveness and School Improvement, 11*(3), 273-303. https://doi.org/10.1076/0924-3453(200009)11:3;1-G;FT273

Ozdamar, K. (2013). *Paket programlar ile istatistiksel veri analizi 1. cilt (Statistical data analysis with package programs, volume 1)*. Eskişehir: Nisan

Pallant, J. (2016). *SPSS survival manual*. The USA: McGraw-Hill Education

Polk, J. A. (2006). Traits of effective teachers. *Arts Education Policy Review, 107*(4), 23-29. https://doi.org/10.3200/AEPR.107.4.23-29

Richardson, P. W., & Watt, H. M. G. (2006). Who chooses teaching and why? Profiling characteristics and motivations across three Australian universities. *Asia-Pacific Journal of Teacher Education, 34*(1), 27-56. https://doi.org/10.1080/13598660500480290

Rots, I., Aelterman, A., Devos, G., & Vlerick, P. (2010). Teacher education and the choice to enter the teaching profession: A prospective study. *Teaching and Teacher Education, 26*(8), 1619-1629. https://doi.org/10.1016/j.tate.2010.06.013

Rowan, B., Chiang, F. S., & Miller, R. J. (1997). Using research on employee's performance to study the effects of teacher on students' achievement. *Sociology of Education, 70*, 256-284. https://doi.org/10.2307/2673267

Saunders, L. (2000). *Effective schooling in rural Africa report 2: Key issues concerning school effectiveness and improvement*. World Bank, Washington, DC. Human Development Network (ERIC Database, ED 453 045). Retrieved from: https://files.eric.ed.gov/fulltext/ED453045.pdf

Senol, S. (2012). *Araştırma ve örnekleme yöntemleri (Research and sampling methods)*. Ankara: Nobel Akademik Yayıncılık.

Shindley Elliott, B. L. (2010). *Effective teacher characteristics: A two nation causal comparative study*. (Doctoral Dissertation). Walden University, Minneapolis

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14. https://doi.org/10.3102/0013189X015002004

Simsek, O. F. (2007). *Yapısal eşitlik modellemesine giriş: Temel ilkeler ve LISREL uygulamaları (Introduction to structural equation modeling: Basic principles and LISREL applications)*. İstanbul: Ekinoks

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62*(4), 339-355. https://doi.org/10.1177/0022487111404241

Stronge, J. H. (2007). *Qualities of effective teachers*. The USA: Association for Supervision and Curriculum Development (ASCD)

Sumbuloglu, V., & Sumbuloglu, K. (2005). *Klinik ve saha araştırmalarında örnekleme yöntemleri ve örneklem büyüklüğü (Sampling methods and sample size in clinical and field research)*. Ankara: Alp Ofset.

Swainston, T. (2008). *A reflective resource for performance management, effective teachers in secondary schools*. London: Network Continuum

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. The USA: Pearson Education

Tschannen Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*(7), 783-805. https://doi.org/10.1016/S0742-051X(01)00036-1

Tschannen Moran, M., & Woolfolk Hoy, A. (2007). The differential antecedents of self-efficacy beliefs of novice and experienced teachers. *Teaching and Teacher Education, 23*(6), 944-956. https://doi.org/10.1016/j.tate.2006.05.003

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning*. The USA: Association for Supervision and Curriculum Development (ASCD)

Vieira A. L. (2011). *Interactive LISREL in practice, getting started with a SIMPLIS Approach*. London: Springer. https://doi.org/10.1007/978-3-642-18044-6

Watt, H. M. G., Richardson, P. W., & Wilkins, K. (2014). Profiles of professional engagement and career development aspirations among USA preservice teachers. *International Journal of Educational Research, 65*, 23-40. https://doi.org/10.1016/j.ijer.2013.09.008

Woolfolk, A. E. (1998). *Educational psychology*. USA: Allyn and Bacon

Yu, Y. (2011). *Pre-service teachers' motivations for choosing a teaching career and intention to teach in urban settings: A multilevel analysis*. (Doctoral Dissertation). Indiana University of Pennsylvania, Pennsylvania

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 436–450

https://doi.org/10.21449/ijate.636370

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Integrating Assessment and Performance Measurement: A Case of an Academic Course for Quality Improvement Actions at a Saudi University

Khalid Mohiuddin[1,*], Mohammad Aminul Islam[1], Shahrear Talukder[2], Mohammed Alghobiri[1], Mohamed Nadhmi Miladi[1], Ahmed Abdelmotlab Ahmed[1]

[1]Management Information Systems, College of Business, King Khalid University, Abha, Saudi Arabia.
[2]Department of English, Bangladesh Army University of Science and Technology, Nilphamari, Bangladesh

ARTICLE HISTORY

Received: Oct. 22, 2019
Revised: May 24, 2020
Accepted: Aug. 09, 2020

KEYWORDS

Course assessment,
Higher education,
Performance indicators,
Rubrics,
Performance evaluation

Abstract: This study aims to assess and measure students' performances using course-key performance indicators (course-KPIs) in an academic course at a Saudi university. The approach includes three aspects of assessment (i) integrating course components and correlating course learning objectives with the program learning domain, (ii) course evaluation using rubrics, and (iii) performance mesurement using a scientific method. Moreover, it presents a novel approach for performance measurement of the course learning skills. In this study, a course has been taken to demonstrate how the KPIs are measured for evaluating students' performances. This approach relies on several specific documents that are developed for the course delivery by following the National Qualification Framework (NQF) in Saudi Arabia and the guidelines of the Accreditation Board for Engineering and Technology (ABET). The performance evaluation outcomes are useful indicators that guide the teachers to improve course learning skills. It also helps the teachers in the quality delivery of the courses and ensures continuous improvement in learning and teaching. This study concludes with an emphasis on the measuring performance using course-KPIs which can be adopted for quality improvement for any academic course in higher education irrespective of data size.

1. INTRODUCTION

Evaluating and measuring students' performance properly in an academic course remains a major concern for higher education institutions (Aoudia et al., 2015). Assessment of an academic practice during a course delivery is one of the essential aspects of students' performance evaluation (University of Technology Sydney, 2010). The course performance evaluation outcomes are useful indicators for the teachers to improve the quality of learning and academic practices. Douglas & Hines (2011) discussed assessment practices, the importance, need, and the complexity of the assessment process in achieving the desired outcomes.

CONTACT: Khalid Mohiuddin ✉ Official email ID: kalden@kku.edu.sa Personal email ID: drkhalidmk70@gmail.com 🖃 Department of Management Information Systems, College of Business, King Khalid University, Abha, Saudi Arabia

ISSN-e: 2148-7456 /© IJATE 2020

To assess students’ performance in an academic course, choosing appropriate assessment types, assessment methods, and assessment activities are very important (Süral, 2016). These assessment components have to be coherent with the course learnings (Light et al., 2009). Further, measuring the students’ performance is equally vital in the quality improvement process (Baeten et al., 2013).

Ideas of the best practice in course evaluation and performance measurement have also started to emerge (Bradley et al., 2015). Kucsera & Svinicki (2010) discussed the academic evaluation, emphasized quality assessment, and considered both systematic and result-orientation in an educational environment. Furthermore, identifying relevant key performance indicators (KPIs) for a course is significant in measuring students’ course learning. An elaborative description of KPI will be found in Section 3.1. These indicators should be coherent to the course objectives and have the potential to measure students’ performance of the course learning (Fernandes et al., 2014).

Indeed, KPIs assist teachers in measuring students’ performance of course learning, and achieving its objectives (Mohiuddin, Rasool, et al., 2019). For course assessment using KPIs, a course teacher needs to determine course performance indicators and define KPIs for measuring students’ performance (Sizer et al., 1992). Moreover, usually KPIs in rubric form (Mohiuddin, Rasool, et al., 2019) describe three levels of performance (see Table 3). Importantly, these KPIs are useful tools to measure course learning in a higher educational environment (Martin & Sauvageot, 2011).

The course-KPIs must be measurable units and significant in measuring course learning performance. KPIs help in judgments and are the authoritative, qualitative, and quantitative measures of key attributes of the functions of an institution (Ramsden, 1991; University of Nottingham Malaysia, 2017). They are viewed as tools that undertake quality assurance, measure the effectiveness and efficiency of the processes to achieve institutional objectives (Bruwer, 1998). Higher education institutions use performance indicators for monitoring academic, institutional performance (Chan, 2015) and internal evaluative procedures (Ramsden, 1991). Brown (2012) suggested that determination of KPIs is one of the primary approaches of performance evaluation in higher education. These indicators are defined considering the course objectives and learning. Further, these KPIs have to be followed while assessing students’ performance of the course learning. The measured performance outcomes help in getting the strengths and weaknesses of the students’ performance. The result of outcomes also helps teachers realize how learning occurs (Pereira et al., 2016). Generally, KPIs evaluate the success of a particular activity in which it engages (Azma, 2011). Based on the performance outcomes, a summary report is prepared that includes the strengths, weaknesses, and suggested actions to improve the course learning. The suggested actions need to be considered in the next cycle of course delivery. Effective implementation of the suggested actions assures continuous improvement of course learning and also helps in achieving the desired learning objectives (Fernandes et al., 2014).

Significantly, the course measure outcomes are the indicators of the students’ course learning. These outcomes precisely show the teachers the strengths and the weaknesses of the course performance and guidelines to improve learning in the next phase of the course delivery. Usually, teachers discuss the measured outcomes with the students after the assessment evaluation and performance measurement. Here, students get the opportunity to know both their performances and course learning abilities, such as their cognitive and interpersonal skills. Further, teachers give more attention to the students’ weak areas, implement the suggested actions while delivering the course at the current and the next cycles. The students need to use the performance measurement results when planning their future course works. Importantly, the whole process guides the teachers and helps the students to improve teaching and learning.

In the institutions with a centralized evaluation system, the primary aim of course assessment is to measure students' performance, analyze the performance outcomes, and implement the suggested actions of the course learning in the next delivery. Some evaluation systems are in place to improve the quality of education (Martin & Sauvageot, 2011). Of the course assessment, the course evaluation outcomes are the end result and useful indicators that should be used wisely for the improvement of course learning, and the academic program the institution offers (Light et al., 2009).

The motivation for this work is the unavailability of a course assessment that integrates course associated components and considers the relationship between the learning domains and course learning. For an academic course, the possible components are Course Specification (CS), Course Report (CR), Program Learning Outcomes (PLOs), Program Specification (PS), learning domains, and course learning skills, and all these are associated coherently. The presented study applies a novel scientific approach to assess students' course leaning and measure their performance. This process also considers the learning domain which is logically corresponding to the course learning. Here, the learning domain, for instance "communication skills" is considered against the course learning. Innovatively, this study describes course-KPIs that are significantly used in measuring students' course learning performance (Mercer-Mapstone & Matthews, 2017).

In the process, we accessed literature available across the top-rated resources. To the best of authors'efforts, we didn't find any study which measures students' performance using course-KPIs (Mercer-Mapstone & Matthews, 2017). This study applies a novel approach to assess students' course learning and measure their performance by applying scientific calculations. This process needs some precise documents which are associated with the course (Klenowski et al., 2006) and useful in evaluating students' course performance. Importantly, the outcomes of the performance measurement correlate with the predefined course learning objectives (CLOs) that determine the intended learning skills of the course (Strydom, 2017). Finally, the measured course outcomes are benchmarked with the outcomes of previous performance measured and new targets are set for the next cycle of course delivery. Indeed, the whole process of measuring course learning improves the quality of learning and teaching (Pereira et al., 2016).

1.1. Contributions

This study introduces an assessment approach that integrates students' assessments in an academic course, its components, and course performance measurement using course-KPIs for continuous improvement. The original contributions of this study are:

- The integrated assessment approach using course-KPIs (Section 3).
- Mapping between course learning, learning domains, and academic program learning outcomes (Table 1).
- A novel approach of course assessment (Section 3.1).
- Measuring students' performances using course-KPIs (Section 3.2).
- The study's approach can be adopted for any academic course and data size.

Apart from these the article comprises of Section 1 which explains the integration of course components and performance measurements in higher education, and the purpose of the study. Section 1.2 describes the study's aim and Section 2 presents the study's adopted methods. Finally, we present the study's findings and conclusions.

1.2. Aim

This study aims to integrate assessment and performance evaluation of students enrolled in an academic course. Significantly, the study also considers course associated components and the logical relation between course learning objectives (CCELT, 2020) with the program learning domain. At the beginning of the course delivery the authors decided to assess students' performance for communication skills of the course learnings and this is in accordance with the "communication skills" of the learning domain because the intended course learning objectives correlate with learning domain "communication skills".

For course assessment, this study measures students' performance using course-KPIs and analysis of the measured outcomes to determine the skills learned from the course contents (Pereira et al., 2016). Here, the assessment methods have to be relative to the course learning (Mohiuddin, Rasool, et al., 2019) and correspond to the learning domain (Mercer-Mapstone & Matthews, 2017). For useful performance measurement, the course-KPIs are defined before the start of the course delivery and agreed on the assessment method. Figure 1 includes the components that are logically associated with the course. Further, based on the students' performance measurement, a summary report is developed that provides the strengths and weaknesses, and the recommended actions (see Table 6) which guide teachers in improving course learning (Pereira et al., 2016). Finally, the result of the performance measured is benchmarked with the previous result and new targets are set for the next performance measurement (see Table 7). This process increases provisions for the teachers in the process of continuous improvement (Strydom, 2017).

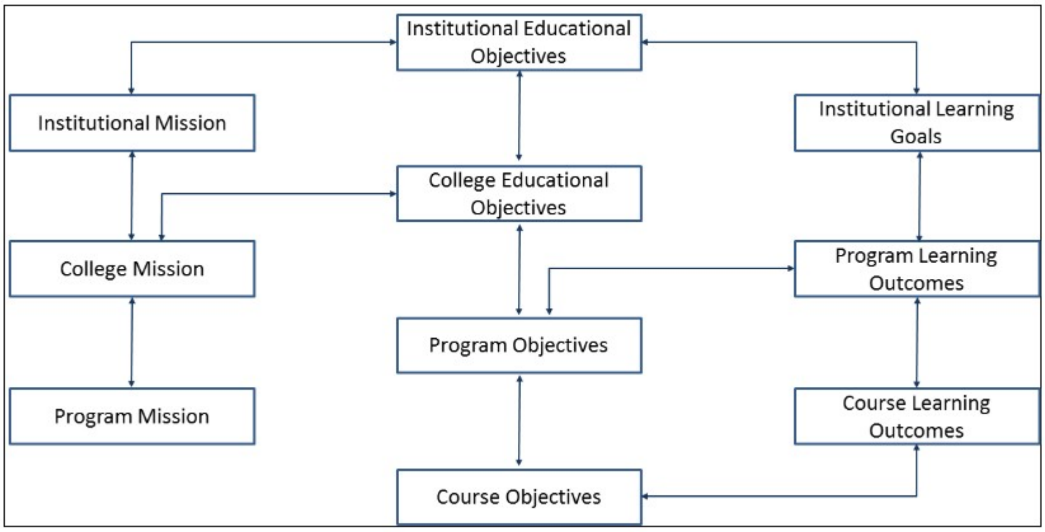

Institutional Educational Objectives
Institutional Mission
Institutional Learning Goals
College Educational Objectives
College Mission
Program Learning Outcomes
Program Objectives
Program Mission
Course Learning Outcomes
Course Objectives

Figure 1. *Coherent components for the performance evaluation of an academic course and a program in higher education.*

Figure 1 describes the components which are associated logically when evaluating students' performance in an academic course of an academic program in higher education. This also shows the importance of other components in achieving course learning objectives. The course objectives are the fundamental learning that students attain at their course completion. The course learning skills and the program learning skills logically map the institutional learning goals, and finally the institutional objectives. These entities are coherent among and aligned together with the mission of the program, college, and institution.

2. METHOD

2.1. Study Context

The study's approach applies to measure students' outcome-based learning performance in an academic course at King Khalid University in Saudi Arabia in the spring semester, 2017- 2018 academic year. The selected three-credit-hour course is a core course and prerequisite to capstone projects of the Information Systems program that is offered in the sixth semester. The

"Seminar" course bears the course code of "MIS492-3." This course's learning outcomes help students to be developed for their capstone projects, specifically in report writing and project presentation. The teaching schedule includes report writing skills of two credits and the oral presentation of one credit. During the course delivery, students participate in both formative and summative assessments. The course teachers measure students' performances against course learning outcomes using the predefined KPIs after each assessment. There are five CLOs of this course split into two groups of "oral presentation/presentation dynamics" and "writing skills."

In this case, students' oral presentation skills of the course content were measured using the defined KPIs. Importantly, course teachers developed these KPIs considering the CLOs approved by the stakeholders such as the academic committee, head of the department, and course coordinator of the department. For the assessment and quality improvement process, some predefined course associated documents were required

This study requires a more in-depth and specific literature review to make it relatable to the study's idea. During the literature review, it is found that various methods have been adopted to evaluate students' performance in academic courses in higher education (Pereira et al., 2016). We have found many studies which evaluate performance using KPIs for both business organization and higher educational institutions. John Sizer (1992) outlined the critical excellence of performance indicators in higher education for achieving desired results. Further, he precisely mentioned the role of performance indicators in higher education and considered them as quality assessment procedures. He also argued for the performance indicators that provide a variety of assessment in the educational system and consider comparative quality judgments. He concluded that political culture, educational funding system and the quality assessment procedures largely impact the role of performance indicators in higher education. Suryadi (2007) developed a framework on key success factors for measuring performance of higher education institutions. KPIs were focused on academic, research and supporting functions. The researcher conducted subjective evaluation using the Analytic Hierarchy Process (AHP) technique. On the other hand, performance indicators were used to measure the teaching performance in Australian Higher Education under a national trail (Ramsden, 1991). Student evaluation was designed using course experience questionnaire which was scored on a 5-point Likert-type scale and several types of analysis were conducted, i.e. item factor analysis, scale internal consistencies and scale validity. However, we couldn't find any study which integrates assessment, course associated components, and course-KPIs that are very useful in measuring students' performance of course learning. From the literature review, it is found that most of the studies are theoretical, and only a few studies are done by collecting the course assessment data. This study bridges the literature lacuna of this neglected area of research (Chan, 2015).

To the best of our efforts and access to the multiple resources (Alstete, 1995; Dawson, 2017; Gibbs, 2003; Haertel, 1999) during our study for course assessment, we couldn't find any specific study that relates our idea (Pereira et al., 2016). We have adopted a systematic approach to evaluate students' performance in the course during the course delivery. The course learning objectives motivate us to evaluate students' performance for the course learning skills along with the program learning skills in the corresponding learning domain. The performance evaluation result helps us to be more specific while delivering the course in the next cycle.

Generally, the maximum limit of students' enrollment for the selected course is twenty. In this case, twenty students were enrolled in the course for the spring semester, 2017. Seven students' enrollment was canceled because of not obeying the initial attendance rules. So, the study sample comprised of only thirteen students that is not big in size. Nevertheless, the study's approach can be used for any number of students and this study also has the potential for quality learning. The course was offered by a department at a Saudi University. All thirteen students participated in the course assessments during the semester. At the beginning of the course delivery, the teachers must consider the previous semester's course evaluation report (e.g., fall 2016), and the benchmarking for the current semester (see Table 7).

2.2. Associated Components

The students' performance data are gathered by the assessment results and the performance evaluation reports. The components which have been shown in Figure 1 contain the associated documents that are considered during course performance evaluation.

Participants: The participants are all the thirteen students, course-teacher, and course coordinator who are directly involved in the evaluation process.

2.3. Associated Documents

For the course delivery, some course-associated documents are required while evaluating and measuring students' performance.

Program specification (PS): It is a precise document that describes the intended learning outcomes of an academic program. It describes program learning skills that are explained through a certain number of PLOs, listed in Table 1. It also describes the curriculum, learning and assessment methods, and other information related to the program (NCAAA, 2012; QAAHE, 2017).

Course specification (CS): It is a prime document that has to be followed while delivering the course. It describes the aims and objectives of the course which covers teaching and assessment methods, and mapping between CLOs and PLOs (Strydom, 2017) and also includes the course evaluation and improvement process. CS should be prepared for all the courses offered in the program and has to be reviewed periodically. Course learning outcomes are statements that describe essential learning that learners have achieved and can reliably demonstrate at the completion of the course (Strydom, 2017). A number of CLOs are written for every course and listed in the CS that have to be mapped with any of the PLOs to correlate with the program learning skills (Sizer et al., 1992).

Course report (CR): A CR is an accumulated document that covers all the activities conducted for the course during the semester. It describes the course execution summary and the evaluation result. It also includes the issues in delivering the course and the suggested actions for the course improvement.

Rubrics: In an academic environment, rubric means a measuring tool used to check students' performance and the quality of their responses. Rubrics usually contain evaluative criteria, quality definitions for those criteria at particular levels of achievement, and a scoring strategy (Popham, 1997). They are often presented in the table format as shown in Table 3 and can be used by teachers when marking and by students when planning their work (Dawson, 2017).

3. ASSESSMENT APPROACH

For assessing students' performance in a course, it is essential to follow the assessment methods and assessment activities described in the course specification. Each course learning outcome listed in the course specification must have logical mapping, at least with one of the program learning outcomes (PLOs). In this case, it is for communication skill (PLO-D), as shown in Table 1. The assessment outcome(s) should be mapped to any of the PLOs, either D1 or D2. It also corresponds to ABET-code-f (ABET, 2017).

Table 1. *Twelve PLOs Distributed into Four Learning Domains*

Learning domains with code	PLO code	Corresponding ABET code
Knowledge (A)	A1, A2, and A3	j, a, e
Cognitive skills (B)	B1, B2, B3, and B4	i, k, c, b
Interpersonal skills (C)	C1, C2, and C3	g, h, d
Communication skills (D)	D1 and D2	f

Table 1 lists the twelve PLOs divided into four learning domains under the national qualification framework (NQF) (NCAAA, 2012). The program learning skills are represented into four learning domains by following the ABET guidelines (ABET, 2017). ABET is an international accreditation agency that accredits academic programs in computing and engineering (Mohiuddin, Islam, et al., 2019). So, we framed three PLOs (A1-A3) in the knowledge domain, (B1-B4) in cognitive skills, (C1-C3) in interpersonal skills, and (D1-D2) assigned for communication skills. Based on the course content, students' evaluation is done for communication skills that correspond to PLO-D2.

Table 2. *Course Evaluation Description Describes at the Beginning of the Course*

PLO code: D2	Level 3: Satisfactory	Level 2: Developing	Level 1: Unsatisfactory
KPI name: Oral presentation delivery	A well-organized oral presentation covering the required contents- delivered effectively.	Covers the required contents but not adequately, missing the technical aspects and flow.	Presentation is organized poorly; neither clear nor the technical aspects of the topic are presented.
Assessment method	Oral presentation		
Assessment activity	Prepare a well-organized presentation on the technical topic from the course content. Demonstrate the technical aspects, the flow, and the presentation mechanics effectively.		
Assessment type	Individual and group – decided by the teacher		

Table 2 describes the assessment framework and the performance indicator (KPI). In the table, the KPI name is 'Oral presentation delivery'. The students' performances have to be evaluated following the assessment method, type, and activity. The performance is assessed in three levels, 'satisfactory', 'developing', and 'unsatisfactory' as shown in Table 3.

3.1. The Process

The course-KPIs are listed in the course specification and should be used for the students' performance measurement. University of Nottingham Malaysia (2017) defined KPIs as the quantitative and qualitative measures used to review the institutional progress against its goals. KPIs' characteristics are realistic, representative, specific, attainable, measurable, and timely. Lord Kelvin (1889) truly said: "If you cannot measure it, you cannot improve it." Moreover, Taticchi et al. (2010) described performance measures as a metric to quantify the efficiency and effectiveness of an action. John Sizer (1992) believed that an indicator represents system performance.

During the course evaluation, measuring students' performance is an essential activity for course improvement (Pereira et al., 2016). The measured outcomes are the critical indicators that guide the teachers for future course delivery (Strydom, 2017). The outcomes provide benificial information which helps its stakeholders in improving the course, the program, and also policy decision making on quality improvement (Dochy et al., 2006).

Use of rubrics: Rubric is a name or heading under which something is classified by comparing particular objectives. Rubrics are developed for several academic activities (Haertel, 1999). Prins et al. (2017) figured the critical use of rubrics, developed it based on the manual of the

American Psychological Association. They suggested that rubrics are effective assessment tools for both teachers and students. They further explained that rubrics are used to make students aware of what is expected, and students get familiar with the grading criteria. Another study (Mohiuddin, Rasool, et al., 2019) conducted a skill-centered assessment in an academic course based on the course-KPIs and rubrics.

Validity and reliability: Generally, the course-KPIs are developed by the teachers with the approval of the knowledge area head in the department. These KPIs and rubrics (see Table 3) are developed and documented well in advance to the course delivery and assessment process. Importantly, the intented course learning outcomes are considered to measure performance, i.e., *validity*, while developing these assessment tools, and they vary with course learning. The measured performance (obtained score) is monitored by the knowledge area head and program coordinator for the consistency and improvement in the course, i.e., *reliability* (Reddy & Andrade, 2010). Every course has its own KPIs and rubrics that are developed, considering both course learning objectives and course learning skills (Martin & Sauvageot, 2011).

Table 3. *Course KPIs in the Form of Rubrics*

PLO-D: An ability to communicate effectively with a range of audiences, from Table 1
PLO (D1): Demonstrate professional competence in written skills
PLO (D2): Communicate verbally with audiences in an effective way

CLOs Vs. D2	Level 3: Satisfactory	Level 2: Developing	Level 1: Unsatisfactory
(KPI-1) Demonstrate understanding of presentation dynamics	Plans and delivers an oral presentation effectively; applies the principle of "tell them" well	Presents key elements of an oral presentation adequately, but "tell them" not clearly applied	Talk is poorly organized; no clear introduction or summary of the talk is presented
(KPI-2) Organize presentation considering audience and time constraint	Presentation has enough detail and appropriate content for the time constraint and the audience	Presentation contains excessive or insufficient detail for the time allowed or level of audience	Presentation is inappropriately short or excessively long; omits key results during the presentation
(KPI-3) Use appropriate technical content	Presentation has appropriate technical content for the time constraint and the audience	Presentation contains excessive or insufficient technical detail for the time allowed or level of audience	Presentation is technically inappropriate; omits key results during a presentation
(KPI-4) Show linguistic command orally	Uses proper American English	Occasionally uses an inappropriate style of English.	Uses poor English
(KPI-5) Illustrate ideas using effective visual aides	Uses visual aids effectively	Visual aides have minor errors or are not always clearly visible	Multiple slides are unclear or incomprehensible

Table 3 represents the sample of five KPIs (1-5) defined and documented for the students' performance measurement in the course. The PLO-D is one of the program learning skills listed in Table 1. D1 and D2 are program learning outcomes split under D. These KPIs correspond to the learning domain-D, i.e., communication skills from Table 1. Each KPI, with its code (1-5), is listed in the first column and is explained into three levels of performance. Level 3 describes the performance standard 'satisfactory', level 2 'developing', and level 1 'unsatisfactory'.

Table 4 represents the sample of a single student's performance in all the KPIs (1-5) in the course assessment. It also shows the performance levels (L), described as 'S-satisfactory,' 'D-developing,' and 'U-unsatisfactory,' and (✓) is the obtained performance by the student. Similarly, the data is collected for all the thirteen students shown in Table 5.

Table 4. *Sample of Single Student's Performance in the Course Assessment*

University ID: 433822625								Course code: ISM492							
Student name: Our student								Course name: Seminar							
Semester: II, Spring 2017								Section number: 1351							
CLOs correspond to PLO-D2															
KPI-code (1-5)	KPI-1			KPI-2			KPI-3			KPI-4			KPI-5		
levels (L)	S	D	U	S	D	U	S	D	U	S	D	U	S	D	U
Obtained	✓				✓		✓				✓		✓		

3.2. Measuring the performance

Measuring the students' performance of the course assessment is possible only when the KPIs are realistic, achievable, and measurable. These students' performances are measured on their attempt in the course assessment. The level of students' performance (1, 2 or 3) decides on students' gain for each KPI, defined in Table 3 and the sample is shown in Table 4.

Table 5. *Students' Performance Measured for KPIs (1-5) for Thirteen Students*

PLO-D2 KPI Nos.	Level 3: (l_3) Satisfactory	Level 2: (l_2) Developing	Level 1: (l_1) Unsatisfactory	N-Total Number	KPI performance out of 5
D2. (1-5)	n_1=03	n_2=04	n_3=06	N=13	2.948

Table 5 shows the average of students' performance data for all the thirteen students who participated in the assessment and the result of KPIs (1-5) measurement. Levels ($l1 - l3$), show the students' performance level in the assessment. These performances are measured using rubrics (from Table 3) and ranked into three groups: satisfactory (n_1=03), developing (n_2=04), and unsatisfactory (n_3=06).

$$KPI = \frac{(n_1 * l_3) + (n_2 * l_2) + (n_3 * l_1)}{(\mathrm{L} * \mathrm{N})} * \mathrm{PS} \tag{1}$$

Where n_1, n_2, n_3, are the three groups of students based on performance
l_1, l_2, l_3, are the three different levels of performance
PS, is the performance scale on 5
L, is the number of performance levels, i.e., 3
N, is the total number of students

$$\mathrm{KPI} = \frac{(3 * 3) + (4 * 2) + (6 * 1)}{(3 * 13)} * 5 \text{ , by applying equation (1)}$$

$$\mathrm{KPI} = \frac{115}{39} = 2.948$$

Table 6 describes the result of students' performance and the teacher's comments. '2.948' is the overall students' performance on scale 5. The evaluator has suggested some actions to be initiated to improve the performance before the next delivery of the course.

Table 6. *The Overall Students' Evaluation Summary Written in Course Report (CR)*

PLO-D - KPI numbers	Level 3: (l_3) Satisfactory	Level 2: (l_2) Developing	Level 1: (l_1) Unsatisfactory	N-Total Numbers	Performance on scale 5
(1-5)	3	4	6	13	2.948

Observations**:**
A few students can organize the presentation correctly.
Some students could not organize the content properly.

Recommendations:
Most of the students should understand the presentation mechanics.
Some of the students should learn the organization of topics in the context.
A few students considerably should learn the presentation skills.

Actions:
Conduct some sessions on presentation skills.
Students should be sent to the language center to improve their communication skills.

3.3. Analysis

The total number of students are categorized into three performance groups from grade points 100:

- Group 'n_1 =03' is graded as 'satisfactory' and their overall share is '23.07%'.
- Group 'n_2 =04' is graded as 'developing' and their overall share is '30.77%'.
- Group 'n_3 =06' is graded as 'unsatisfactory' and their overall share is '46.15%'.

Meaningful outcomes: The teachers translate the obtained numerical values into meaningful outcomes by following Table 3 and 6. The same result is shown graphically in Figure 2.

- '23.07%' students have presented the concept, organized presentation in the context, and gave presentation convincingly.
- '30.77%' students have covered the topic but not in an adequate way, and the presentation was not convincing.
- '46.15%' students were unable to organize the presentation in the context. Even the concept in the topic was not clear.

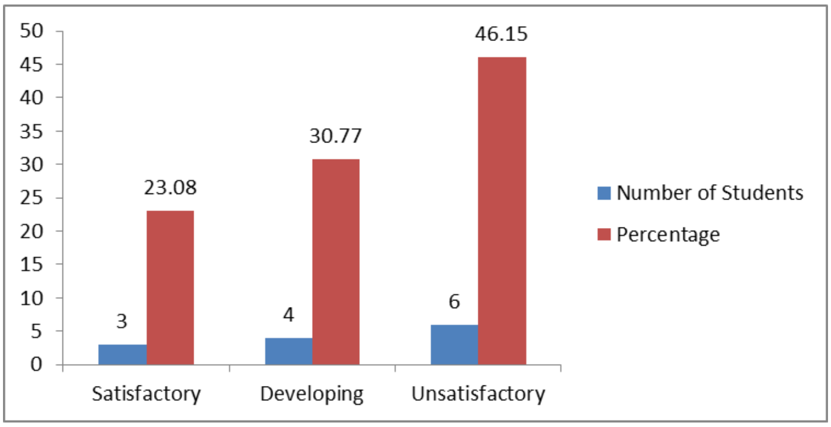


Figure 2. *Students' performance measured using KPIs out of 100 grade points.*

3.4. Course Improvement Plan

The students' performance records are kept into a formal document called 'course report-CR'. This CR is the executed summary of the course specification. The teachers preserve the assessments' record for the future practices. Based on the performance outcomes, a course quality improvement plan develops. The course coordinator monitors the assessments' process and checks the possibilities to implement the suggested actions before the next course delivery.

Effective implementation of the suggested actions assures quality improvement in the course delivery and learning.

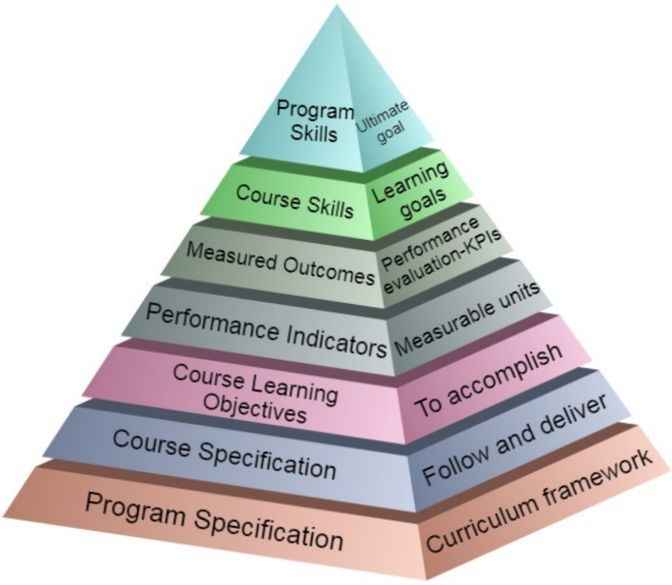


Figure 3. *The bottom-up approach and associated entities of an academic course in higher education*.

Figure 3 shows the coherent entities that need to be considered when evaluating students' performance in an academic course. The process begins referring to the program specification, executing course specifications, and ends-up when the students' performance is measured. During the process, these coherent entities must be followed to achieve all the learning skills. The evaluation result indicates that the skills are learned by the students and demonstrates the accomplishment of program skills. The evaluation process is implemented for every single course offered from the curriculum of the academic program. Surely, this achieves both learning objectives and quality teaching.

3.5. Benchmarking

Benchmarking has been emerged as a useful tool for staying competitive (Alstete, 1995). The strategy of benchmarking is significant and is being used as an instructional model in academic institutions to improve quality (Alstete, 1995). Stakeholders in higher education have realized the increasing importance of benchmarking for continuous improvement.

In this approach, the performance evaluation result is benchmarked with the previous performance result of the course. This will increase the option of implementing a course improvement plan before the next course delivery. Also, this helps in the process of continuous improvement of course learning and teaching.

Table 7. *Benchmarking the Overall Students' Performance Measured in the Course*

CLOs map to PLO-D (An ability to communicate effectively with a range of audiences)	
Assessment year**:** spring semester, 2017. Course learning skills map to PLO-D1 and D2, as shown in Table 3.	
Learning domain	Communication skill
Target benchmark	3.5 – was set for spring semester, 2017
Measured performance	2.948– is achieved for spring semester, 2017
Internal benchmark	3.5 – was set by the authority for spring semester, 2017
External benchmark	4.0 and above, set by external advisory board
New target benchmark	3.25 – set by the coordinator for fall semester, 2017

Table 7 describes the benchmarking of the students' performance measurement of the learning skills in the course assessment. The first column shows different benchmarking fields and the corresponding values in the second column.

3.6. Key issues

Klenowski et al. (2006) highlighted the issues of effective learning and portfolio used in higher education and Pedrosa de Jesus (2009) explained the assessment methods, issues that were aligned with teaching and learning. Our study follows some specific documents that describe activities to be followed for the course delivery. The main issue is to develop these documents during the course delivery under NQF (ETEC, 2018) in Saudi Arabia. Subsequently, maintaining the course evaluation reports is critical for the stakeholders in every semester. Further, the study follows ABET guidelines for measuring the skills learned from the course content (ABET, 2017). The other key issue is that the process has to be followed more effectively on each cycle of the course delivery.

4. CONCLUSION

This study facilitates to minimize the gap between the unavailability of good number of research and useful study on measuring student performance using course-KPIs in higher education. This process helps in achieving course learning objectives and program learning skills that are the prime aspects of this study. Skilled graduates can be produced by measuring the students' performance of the courses offered in the academic program.

Most of the existing assessment methods are course learning oriented. This study demonstrates a novel approach to measure and assess students' course learning skills by applying course-KPIs exemplifying scientific calculations. Teachers will find those calculations convenient to measure their learners' achievement. Though this study uses KPIs and relevant rubrics to show a single skill namely 'Communication Skills,' other skills surely can be modeled upon this KPIs to measure effectively the skills performance.

Moreover, the performance evaluation result indicates the strengths and weaknesses of the students' performance. This result also guides the teachers in preparing course improvement plan on continuous improvement. This study presents the diagraming among course learning, learning domains, and academic program learning outcomes keeping KPIs at the center to guide the whole assessment and evaluation process. It also helps in identifying how to measure students' performance using rubrics. The correlation between course evaluation and its components, and benchmarking are the great significance of this study. Indeed, this provides a suitable direction to the teachers for quality teaching. The approach demonstrated in this study can be adopted for any academic course in higher education irrespective of the number of participating students.

Finally, the presented study assures that the quality improvement in teaching-learning by following the approach will be enhanced. Future research with this approach is undoubtedly adaptable for any academic course with different sample sizes in higher education. Indeed, by following this approach, the study's validity and effectiveness can be compared in any state or region. Significantly, educators get a fair picture by practicing this assessment approach and a significant change in the quality improvement process.

Declaration of Conflicting Interests and Ethics

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

ORCID

Khalid MOHIUDDIN https://orcid.org/0000-0001-7531-4512
Mohammad Aminul ISLAM https://orcid.org/0000-0001-8269-2394
Shahrear TALUKDER https://orcid.org/0000-0002-9840-4139
Mohammed ALGHOBIRI https://orcid.org/0000-0002-6414-739X
Mohamed Nadhmi MILADI https://orcid.org/0000-0002-9862-0034
Ahmed Abdelmotlab AHMED https://orcid.org/0000-0002-4363-8261

5. REFERENCES

ABET (2017). Criteria for Accrediting Computing Programs, 2016 2017. *ABET*. https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-computing-programs-2016-2017/

Alstete, J. W. (1995). Benchmarking in Higher Education: Adapting Best Practices to Improve Quality. *ERIC Development Team*. http://files.eric.ed.gov/fulltext/ED402800.pdf

Aoudia, M., Marji, K., & AlQahsi, D. A.-D. (2015). Assessment of Higher Education Quality by Using Cohort of First-year in University. *Procedia - Social and Behavioral Sciences*, *191*, 330–335. https://doi.org/10.1016/j.sbspro.2015.04.310

Azma, F. (2011). The Quality Indicators of Information Technology in Higher Education. *Procedia - Social and Behavioral Sciences*, *30*, 2535-2537. https://doi.org/10.1016/j.sbspro.2011.10.494

Baeten, M., Struyven, K., & Dochy, F. (2013). Student-centred Teaching Methods: Can They Optimise Students' Approaches to Learning in Professional Higher Education? *Studies in Educational Evaluation*, *39*(1), 14–22. https://doi.org/10.1016/j.stueduc.2012.11.001

Bradley, K. D., Snyder, E. M., & Tombari, A. K. (2015). Higher Education End-of-Course Evaluations: Assessing the Psychometric Properties Utilizing Exploratory Factor Analysis and Rasch Modeling Approaches. *International Journal of Assessment Tools in Education*, *3*(1). https://ijate.net/index.php/ijate/article/view/90

Brown, C. (2012). Application of the Balanced Scorecard in Higher Education Opportunities and Challenges. *Society for College and University Planning*, *40*(4), 40–50.

Bruwer, J. (1998). First Destination Graduate Employment as Key Performance Indicator: Outcomes Assessment Perspectives. *Journal of Institutional Research in Australasia*, *8*(2), 61–91. http://www.aair.org.au/app/webroot/media/pdf/AAIR Fora/Forum1998/Bruwer.pdf

CCELT. (2020). *Writing Course Goals/Learning Outcomes and Measurable Learning Objectives*. Iowa State University. https://www.celt.iastate.edu/teaching/preparing-to-teach/tips-on-writing-course-goalslearning-outcomes-and-measur

Chan, V. (2015). Implications of Key Performance Indicator Issues in Ontario Universities Explored. *Journal of Higher Education Policy and Management*, *37*(1), 41–51. https://doi.org/10.1080/1360080X.2014.991531

Dawson, P. (2017). Assessment Rubrics: Towards Clearer and More Replicable Design, Research and Practice. *Assessment & Evaluation in Higher Education*, *42*(3), 347–360. https://doi.org/10.1080/02602938.2015.1111294

Dochy, F., Segers, M., & Sluijsmans, D. (2006). The Use of Self-, Peer and Co-assessment in Higher Education: A Review. *Studies in Higher Education*, *24*(3), 331–350. https://doi.org/10.1080/03075079912331379935

Douglas, E. M., & Hines, D. A. (2011). The Helpseeking Experiences of Men Who Sustain Intimate Partner Violence: An Overlooked Population and Implications for Practice. *Journal of Family Violence*, *26*(6), 473–485. https://doi.org/10.1007/s10896-011-9382-4

ETEC. (2018). *National Framework for Public Education Curricula Standards*. Education and Training Evaluation Commission. https://etec.gov.sa/en/productsandservices/NCSEE/Cevaluation/Pages/NATIONALFRAMEWORK-.aspx

Fernandes, S., Mesquita, D., Flores, M. A., & Lima, R. M. (2014). Engaging Students in Learning: Findings from a Study of Project-led Education. *European Journal of Engineering Education*, *39*(1), 55–67. https://doi.org/10.1080/03043797.2013.833170

Gibbs, C. (2003). Explaining Effective Teaching: Self-efficacy and Thought Control of Action. *The Journal of Educational Enquiry*, *4*(2). https://www.ojs.unisa.edu.au/index.php/EDEQ/article/view/520/0

Haertel, E. H. (1999). Performance Assessment and Education Reform. *Phi Delta Kappan*, *80*(9), 662. https://www.questia.com/library/journal/1G1-54618911/performance-assessment-and-education-reform

Klenowski, V., Askew, S., & Carnell, E. (2006). Portfolios for Learning, Assessment and Professional Development in Higher Education. *Assessment & Evaluation in Higher Education*, *31*(3), 267–286. https://doi.org/10.1080/02602930500352816

Kucsera, J. V., & Svinicki, M. (2010). Rigorous Evaluations of Faculty Development Programs. *The Journal of Faculty Development*, *24*(2), 5. https://eric.ed.gov/?id=EJ897466

Light, G., Cox, R., & Calkins, S. (2009). *Learning and Teaching in Higher Education : The Reflective Professional*. Sage Publications Ltd. https://books.google.com.sa/books?id=BDtdBAAAQBAJ&source=gbs_book_other_versions

Martin, M., & Sauvageot, C. (2011). Constructing an Indicator System or Scorecard for Higher Education: A Practical Guide. In *Intenational Institute for Educational Planning*. http://uis.unesco.org/sites/default/files/documents/constructing-an-indicator-system-or-scorecard-for-higher-education-a-practical-guide-2011-en.pdf

Mercer-Mapstone, L. D., & Matthews, K. E. (2017). Student Perceptions of Communication Skills in Undergraduate Science at an Australian Research-intensive University. *Assessment and Evaluation in Higher Education*, *42*(1), 98-114. https://doi.org/10.1080/02602938.2015.1084492

Mohiuddin, K., Islam, A., Mohd, S., & Shariff, M. (2019). Evaluation of an Academic Program: The Case of Computing Accreditation Commission Framework in Higher Education. *International Journal of Emerging Technologies in Learning*, *14*(11), 70–91. https://online-journals.org/index.php/i-jet/article/view/10178/5719

Mohiuddin, K., Rasool, A. M., Mohd, M. S., & Mohammad, R. H. (2019). Skill-Centered Assessment in an Academic Course: A Formative Approach to Evaluate Student Performance and Make Continuous Quality Improvements in Pedagogy. *International Journal of Emerging Technologies in Learning*, *14*(11), 92–106. https://online-journals.org/index.php/i-jet/article/view/10275/5720

NCAAA, S. A. (2012). National Commission for Academic Accreditation and Assessment Handbook for Quality Assurance and Accreditation in Saudi Arabia, Part 1-The System for Quality Assurance and Accreditation. *NCAAA*. http://www.kfupm.edu.sa/deanships/dad/Documents/AAC/NCAAA Documents/H1. Handbook Part 1.pdf

Pedrosa de Jesus, H. (2009). The Role of Students' Questions in Aligning Teaching, Learning and Assessment: A Case Study from Undergraduate Sciences. *Assessment & Evaluation in Higher Education*, *34*, 193–208. https://doi.org/10.1080/02602930801955952

Pereira, D., Flores, M. A., & Niklasson, L. (2016). Assessment Revisited: A Review of Research in Assessment and Evaluation in Higher Education. *Assessment & Evaluation in Higher Education*, *41*(7), 1008-1032. https://doi.org/10.1080/02602938.2015.105523

3

Popham, W. J. (1997, October). What's Wrong—and What's Right—with Rubrics. *Educational Leadership*, 72-75. http://www.ascd.org/publications/educational-leadership/oct97/vol55/num02/What's-Wrong—and-What's-Right—with-Rubrics.aspx

Prins, F. J., de Kleijn, R., & Tartwijk, J. van. (2017). Students' Use of a Rubric for Research Theses. *Assessment & Evaluation in Higher Education*, *42*(1), 128–150. https://doi.org/10.1080/02602938.2015.1085954

QAAHE. (2017). Guidelines for Preparing Programme Specifications. In *The Quality Assurance Agency for Higher Education*. https://www.tsu.ge/data/file_db/qa_docs/guidelines for programme specifications.pdf

Ramsden, P. (1991). A Performance Indicator of Teaching Quality in Higher Education: The Course Experience Questionnaire. *Studies in Higher Education*, *16*(2), 129–150. https://doi.org/10.1080/03075079112331382944

Reddy, Y. M., & Andrade, H. (2010). A Review of Rubric Use in Higher Education. *Assessment & Evaluation in Higher Education*, *35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Sizer, J., Spee, A., & Bormans, R. (1992). The Rôle of Performance Indicators in Higher Education. *Springier-Higher Education*, *24*(2), 133-135. https://link.springer.com/content/pdf/10.1007/BF00129438.pdf

Strydom, F. (2017). Higher Education Learning Outcomes Assessment: International Perspectives. *Assessment & Evaluation in Higher Education*, *42*(3), 492–494. https://doi.org/10.1080/02602938.2016.1139097

Süral, S. (2017). The Development Study of Thoughts Scale Towards Measurement and Assessment Course on High Education. *International Journal of Assessment Tools in Education*, *4*(1), 79-95. https://dergipark.org.tr/en/pub/ijate/issue/23899/270300

Suryadi, K. (2007). Framework of Measuring Key Performance Indicators for Decision Support in Higher Education Institution. In *Journal of Applied Sciences Research,* 3(12), 1689-1695

Taticchi, P., Tonelli, F., & Cagnazzo, L. (2010). Performance Measurement and Management: A Literature Review and a Research Agenda. *Measuring Business Excellence*, *14*(1), 4–18. https://doi.org/10.1108/13683041011027418

University of Nottingham Malaysia. (2017). *Key Performance Indicators*. University of Nottingham Malaysia. https://www.coursehero.com/file/p18pg87/PROS-1010-Guideline-3-Key-Performance-Indicators-State-of-Victoria-2010-Version/

University of Technology Sydney. (2010). Seven Propositions for Assessment Reform in Higher Education. In *Australian Learning & Teaching Council*. https://www.uts.edu.au/sites/default/files/Assessment-2020_propositions_final.pdf

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 451–487
https://doi.org/10.21449/ijate.660353
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

Comparison of Confirmatory Factor Analysis Estimation Methods on Binary Data

Abdullah Faruk Kilic [1], **Ibrahim Uysal** [2,*], **Burcu Atar** [3]

[1]Department of Educational Sciences, Faculty of Education, Adıyaman University, Adıyaman, Turkey
[2]Department of Educational Sciences, Faculty of Education, Bolu Abant İzzet Baysal University, Bolu, Turkey
[3]Department of Educational Sciences, Faculty of Education, Hacettepe University, Ankara, Turkey

ARTICLE HISTORY

Received: Dec. 17, 2019
Revised: July 01, 2020
Accepted: Aug. 09, 2020

KEYWORDS

Confirmatory factor analysis,
Estimation methods,
Binary data,
Simulation

Abstract: This Monte Carlo simulation study aimed to investigate confirmatory factor analysis (CFA) estimation methods under different conditions, such as sample size, distribution of indicators, test length, average factor loading, and factor structure. Binary data were generated to compare the performance of maximum likelihood (ML), mean and variance adjusted unweighted least squares (ULSMV), mean and variance adjusted weighted least squares (WLSMV), and Bayesian estimators. As a result of the study, it was revealed that increased average factor loading and sample size had a positive effect on the performance of the estimation methods. According to the research findings, it can be said that the methods are sufficient to estimate average factor loading and interfactor correlations, regardless of the estimation methods, in most of the conditions where the average factor loading is 0.7. In small sample sizes particularly, the interfactor correlation was underestimated for skewed indicator conditions. According to the findings of the study, although there is not the most accurate method in all conditions, it can be recommended to use ULSMV method because it performs adequately in more conditions.

1. INTRODUCTION

Most researchers conducting research in social, behavioral, and educational sciences usually work on psychological attributes. Psychological attributes, also named as constructs, are theoretical concepts. Psychological constructs cannot be directly observed: the degree to which a construct characterizes an individual can only be predicted by observing the behaviors of the individual (Crocker & Algina, 2008). To analyze the relationships among observed variables and latent constructs, researchers widely use structural equation modeling techniques (Byrne, 2016; Raykov & Marcoulides, 2006). The use of confirmatory factor analysis (CFA) is also widely accepted as one of the structural equation models to examine the construct validity of the hypothesis (AERA et al., 2014).

When the scale development and adaptation studies in the literature are examined, it is observed that CFA is frequently used for collecting evidence for construct validity. Acar-Güvendir and

CONTACT: İbrahim UYSAL ✉ ibrahimuysal06@gmail.com Department of Educational Sciences, Faculty of Education, Bolu Abant İzzet Baysal University, Bolu, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

Özer-Özkan (2015) and Şahin and Boztunç Öztürk (2018) examined scale development studies and they reported that CFA was used in 61% and 52% of these studies, respectively. Deciding the estimation method used in CFA is all-important to obtain unbiased parameter estimations. For this reason, it is also important to examine which estimation method is unbiased.

When the literature is examined, there are many studies comparing CFA estimation methods. One of the most comprehensive of these studies is one conducted by Forero et al. (2009). In this study, the researchers studied 324 simulation conditions and the performance of diagonally weighted least squares (DWLS) and unweighted least squares (ULS) estimation methods was compared in terms of sample size, measurement model, test lengths, factor loadings, and categories of indicators. As a result of the research, it was reported that both methods had similar results but ULS had more accurate and less variable results for parameter estimations.

Another comprehensive study in the literature was conducted by Flora and Curran (2004). In this study, they manipulated latent response (y*) distributions, model specifications, sample sizes, and number of categories (160 simulation conditions). As a result of the study, it was reported that, while WLS requires a large sample size, robust WLS performs better for all conditions. Also, they reported that polychoric correlation is strong against moderate violations of normality.

The study conducted by Rhemtulla et al. (2012) aimed to compare the performance of robust ML and robust categorical least squares estimation (cat-LS) method. CFA model size, underlying distribution, number of indicator categories, threshold symmetry, and sample size were manipulated. As a result of the study, it was reported that ML was more sensitive to asymmetric thresholds. The cat-LS method was suggested for indicators which have fewer than five categories.

When the other studies in the literature were researched, it was observed that there are many studies which examine datasets consists of five categories indicators (Babakus et al., 1987; Lei, 2009; Morata-Ramirez & Holgado-Tello, 2013; B. O. Muthén & Kaplan, 1985; Potthast, 1993). There are also studies examining data consisting of other than five categories. Dolan (1994) used 2, 3, 5 and 7 categories indicators, for example; Green et al. (1997) used datasets with 2, 4, and 6 categories and continuous indicators. Flora and Curran (2004) used 2 and 5 categories indicators; Beauducel and Herzberg (2006) used 2, 3, 4, 5, and 6 categories indicators; Forero et al. (2009) used 2 and 5 categories indicators; Yang-Wallentin et al. (2010) used 2, 5 and 7 categories indicators; Rhemtulla et al. (2012) used 2, 3, 4, 5, 6, and 7 categories indicators; Liang and Yang (2014) used 2 and 4 categories indicators, and Moshagen and Musch (2014) used 2 and 5 categories indicators. However, in most of these studies, frequentist estimation methods have been compared. In addition, in most of these studies, datasets were generated such that the factor loadings of the indicators were equal (Beauducel & Herzberg, 2006; Flora & Curran, 2004; Forero et al., 2009; Liang & Yang, 2014; B. O. Muthén & Kaplan, 1985; Nestler, 2013; Shi et al., 2018). However, in real life applications, it is difficult to have equal factor loadings of all indicators.

The motivation of the present study is to compare Bayesian estimation method and frequentist estimation methods under simulation conditions. Studies comparing the performance of Bayesian method with frequentist methods are also found in the literature. For example, Liang and Yang (2014) compared the performance of WLSMV and Bayesian (informative priors and non-informative priors) methods in 96 simulation conditions. However, in this study, the factor loadings of the indicators were manipulated to be 0.50 and 0.80. There was also no comparison with the performance of the ML method used by default in most software. Xu (2019) compared robust ML (MLR) and Bayesian estimation method's performance in 27 simulation conditions. However, in this study, factor loadings of all indicators were fixed as 0.7. In the study conducted by Önen (2019), ML and Bayesian estimation methods were compared in terms of detecting model misspecification, using 0.30 and 0.80 as factor loadings in the simulation study.

When the researches in the literature were studied, although many datasets consisting of different categories of indicators were used, we could not find any paper which compared ML, ULSMV, WLSMV, and Bayesian estimation methods and, at the same time, not have fixed factor loadings of indicators. The differentiation of factor loadings of the indicators would be more appropriate for real situations. Therefore, in this study, the average factor loading, which is more suitable for real conditions, was determined as the simulation condition and the factor loadings of the indicators were generated to be different from each other (details are given in the method section). At the same time, the aim is to compare the performance of Bayesian and frequentist estimation methods (ML, ULSMV and WLSMV) for binary indicators. The present study differs from other studies in the literature in terms of compared estimation methods (frequentist vs. Bayesian) and not fixing the factor loadings of all indicators. Therefore, it is considered that the current research will contribute to the literature in order to examine which estimation methods (ML, ULSMV, WLSMV, or Bayes) perform better in binary data under different simulation conditions and will help researchers in practice.

2. METHOD

This research was designed as a Monte Carlo simulation study. Monte Carlo simulations use random sampling for a statistical model across varying conditions (Harrison, 2010). Thus, suggestions can be made by investigating the effects of different factors for the statistical model (Gilbert, 1999). The main purpose of this study was to investigate estimation methods performance under different simulation conditions. For this purpose, unlike other studies, average factor loading is considered as a simulation condition. In addition, the performance of Bayesian and frequentist estimation methods.

2.1. Estimation Methods

Estimation methods differ from each other in terms of the analysis processes they use and assumptions. In general, there are four types of estimation methods: maximum likelihood; unweighted least squares; generalized least squares, and asymptotically distribution-free (generally called as weighted least squares). Each estimation approach tries to minimize the corresponding fit function (Raykov & Marcoulides, 2006). The fit function expresses the fit between the covariances obtained from the sample and the covariances obtained from the model established by the researcher (Kline, 2016).

The maximum likelihood (ML) method is the most commonly used estimation method by researchers (Bollen, 1989). Being the default estimation method in most software, ML may be used more frequently in research. ML estimation method assumes that indicators are measured on continuous scales and requires a large sample size. Although ML requires a continuous data set, it was seen that ML is used for binary data sets in the literature. For example, ML method was used in binary data sets in Koğar and Yılmaz Koğar's (2015) study. The ULS method makes estimations under the assumption of continuous variables holding multivariate normal distribution. In addition, all variables in this process should take place on the same scale (Kline, 2016). The weighted least squares (WLS) method has no distributional assumption. Estimations can be made for both continuous and categorical indicators. However, this estimation method needs a large sample size (Kline, 2016). Bayesian methods differs from other frequency-based methods in terms of fixed and free parameters. When the ML method calculates the values which will make the obtained likelihood function maximum, Bayesian methods make estimations by combining the prior distribution of the data with the posterior distribution of parameter estimation (B. O. Muthén & Asparouhov, 2012).

The estimation methods explained above are considered as essential methods; modified estimation methods have been obtained with the help of some corrections via essential methods.

In weighted least square parameter estimates using a diagonal weight matrix with standard errors and mean adjusted chi-square test statistic (WLSM) method, which is developed based on the WLS estimation method, the average corrected chi-square test statistics are produced by using full weight matrix. When the variances in the WLSM method are corrected, a modification of the WLS method is obtained with the WLSMV method. Full weighted matrix is used in the WLSMV method (L. K. Muthén & Muthén, 2015). In ULS parameter estimates with standard errors and a mean and variance adjusted (ULSMV) method, which is developed based on the ULS method, both the averages and the variances are corrected and the chi-square test statistics is calculated over the full weighted matrix (L. K. Muthén & Muthén, 2015).

2.2. Simulation Design

Five factors were manipulated in this simulation study: *(i)* sample size (200, 500 and 1,000); *(ii)* distribution of indicators (left skewed, normal and right skewed); *(iii)* test length (10 and 20 indicators); *(iv)* average factor loading (0.4 and 0.7), and *(v)* factor structure (unidimensional, two factors [φ = 0], two factors [φ = 0.3], two factors [φ = 0.6]). Full crossed design was adopted for simulation conditions.

The sample sizes were 200, 500 and 1,000. Boomsma (1985) suggests a sample size of at least 200 to avoid non-convergence and improper solutions. In addition, Mulaik (2009) states that a sample size of less than 200 is inadequate for statistical inference purposes with chi-square statistics in CFA. Liang and Yang (2014) also emphasize that there are very few studies with sample sizes of less than 200. Therefore, 200 was specified as a minimum sample size. Other sample sizes were specified to examine the effects of sample sizes on the performance of the estimation methods. In addition, the 1,000 sample size was included in the study as recommended as a minimum sample size by some researchers (Comrey & Lee, 1992; Floyd & Widaman, 1995; Gorsuch, 1974; Guadagnoli & Velicer, 1988; Streiner, 1994).

The distribution of indicators was manipulated to be left-skewed, normal and right-skewed. The ML estimation method estimates parameters under the assumption that the variables meets multivariate normal distribution (Tabachnik & Fidell, 2012). WLS and ULS are asymptotic distribution free methods (Brown, 2015). The aim was to examine the estimation methods performance when indicators were skewed. Therefore, the skewness of the indicators was specified as a simulation condition (details are in the data generation section).

The test length conditions were manipulated as 10 and 20 indicators. In order to examine the performance of the estimation methods in short tests, a test length of 10 indicators was specified as a simulation condition. The 20 item condition was determined to examine how the results change when the test length increases.

The average factor loading was specified as 0.4 (low) and 0.7 (high). Unlike other studies, the factor loadings of the indicators were generated to be different from each other. Since the lowest factor loading was suggested to be 0.4 (Stevens, 2009) or .32 (Tabachnik & Fidell, 2012), the average factor loading was specified as 0.4 for low factor loading. Since the average factor loading was used in current study, the condition of 0.40 was added as the lowest average factor loading. Because the factor loadings of the items can be smaller than 0.40 (see Table 2). As we aimed to investigate the performance of estimation methods at high factor loading, the average factor loading was specified as 0.7.

Factor structure is considered as unidimensional and two-factors (φ = 0, 0.3 and 0.6). When the studies in the literature are examined, it is observed that achievement tests are usually unidimensional (Anıl et al., 2010; Kılıç & Kelecioğlu, 2016) but, in some cases, two-factors structures may also occur (Lissitz et al., 2012; Thissen et al., 1994). Therefore, both unidimensional and two-factors structures were specified as simulation conditions. Interfactor correlations were set to φ = 0, 0.3 and 0.6 and were manipulated to examine how the magnitude of the relationship between factors in two-factors structures affected the performance of estimation methods. While φ = 0 was specified because of the performance of the estimation methods in unrelated structures, φ = 0.3 was specified because of its frequent use in studies (Curran et al., 1996; Flora & Curran, 2004; Li, 2016). Thus, the results of the study can be compared to other studies in the literature. φ = 0.6 was specified because it offered the chance of examining how the increase of interfactor correlation affected the performance of the estimation methods. Thus, the aim is to examine the performance of estimation methods under these changing conditions. Table 1 contains a summary of the factors held constant and manipulated factors with their levels.

Table 1. *Simulation conditions*

Fixed Factor		Manipulated Factors			
Number of Categories of Indicators	Sample Size	Distribution of Indicators	Test Length	Average Factor Loading	Factor Structure (Model)
1-0	200 500 1,000	Left-Skewed Normal Right-Skewed	10 20	0.40 0.70	Unidimensional Two Factors (φ = 0) Two Factors (φ = 0.3) Two Factors (φ = 0.6)

Full crossed factorial design was used in the study. By crossing each condition, 3x3x2x2x4=72 simulation conditions have been studied. The number of indicators is equally divided between the factors in two-factorial models. For example, in two-factors models with 10 indicators, five indicators were included in each factor. For each condition, 1,000 replications were obtained. The models examined in the study are presented in Figure 1.

The factor loadings were specified as Table 2.

Table 2. *Factor loadings used in study*

Item Number	Figure 1.a Factor Loadings		Figure 1.c Factor Loadings		Figure 1.b Factor Loadings		Figure 1.d Factor Loadings	
	Average Factor Loading							
	0.4	0.7	0.4	0.7	0.4	0.7	0.4	0.7
1	0.39	0.68	0.39	0.72	0.36	0.68	0.37	0.68
2	0.37	0.72	0.37	0.73	0.40	0.73	0.37	0.73
3	0.38	0.68	0.38	0.69	0.40	0.71	0.38	0.71
4	0.39	0.68	0.39	0.68	0.39	0.69	0.44	0.69
5	0.45	0.7	0.45	0.70	0.43	0.72	0.34	0.72
6	0.39	0.72	0.39	0.73	0.39	0.69	0.35	0.69
7	0.42	0.7	0.42	0.72	0.40	0.70	0.45	0.70
8	0.43	0.73	0.43	0.69	0.39	0.71	0.44	0.71
9	0.42	0.72	0.42	0.71	0.40	0.73	0.36	0.73
10	0.34	0.70	0.34	0.67	0.44	0.69	0.46	0.69
11					0.38	0.72	0.36	0.70
12					0.42	0.72	0.40	0.68
13					0.41	0.71	0.40	0.71
14					0.39	0.7	0.39	0.72
15					0.45	0.71	0.43	0.70
16					0.38	0.72	0.39	0.70
17					0.42	0.71	0.40	0.67
18					0.41	0.71	0.39	0.74
19					0.41	0.71	0.40	0.68
20					0.40	0.72	0.44	0.71

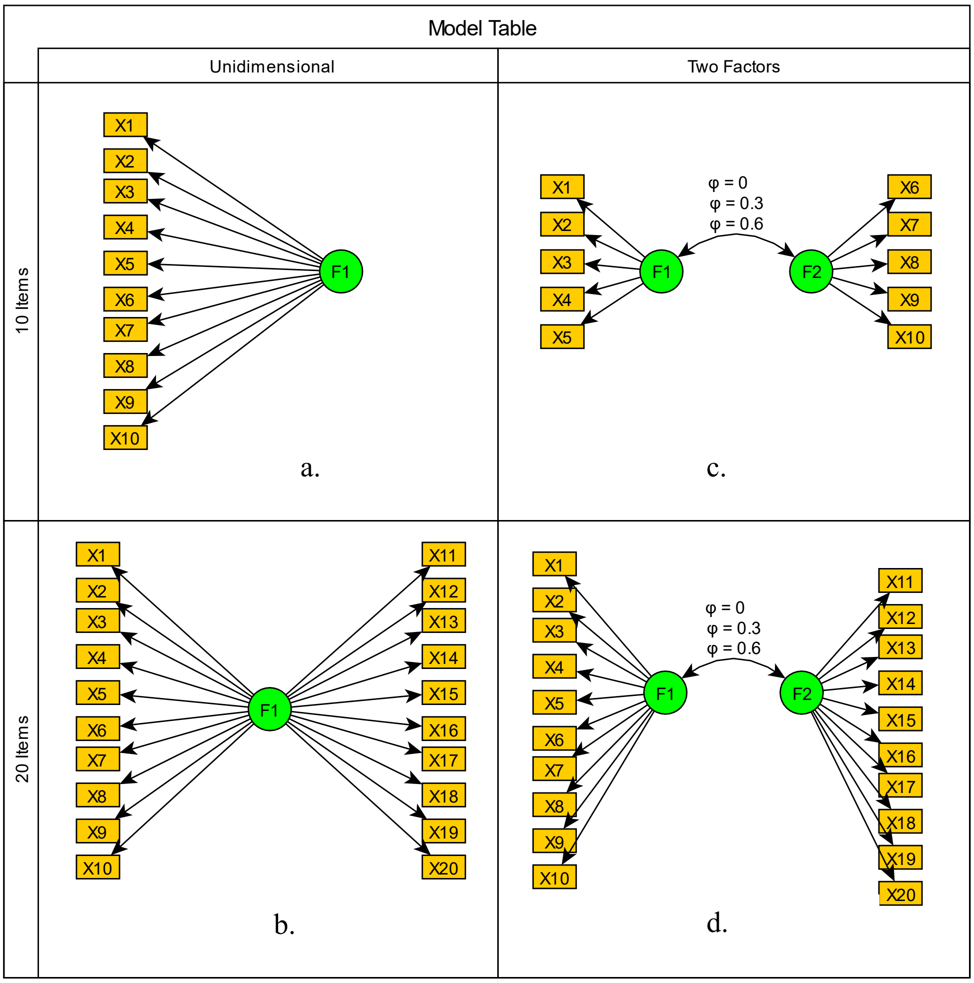


Figure 1. *Models examined in the research*

2.3. Data Generation

A latent response variable framework was used in data generation (Brown, 2015; B. O. Muthén & Asparouhov, 2002). Accordingly, the datasets were firstly generated as continuous which holds multivariate normal distribution. Then, the datasets were categorized according to the skewness of the indicators using threshold values. The threshold values are specified as {0} for normal distribution, {1.05} for right-skewed and {-1.05} for left-skewed. In this case, the mean skewness values of the indicators are 0 for normal distribution, 2.00 for right-skewed distribution and -2.00 for left-skewed distribution. Kurtosis values are 2, 5 and 5 respectively. The lavaan package (Rosseel, 2012) in the R software (R Core Team, 2018) was used for data generation.

2.4. Data Analysis

The Mplus software (L. K. Muthén & Muthén, 2012) was used to analyze the generated data. The MplusAutomation package (Hallquist & Wiley, 2017) was used to analyze the simulated data and to obtain the outputs of the analyses. The performance of ML, mean and variance adjusted weighted least squares (WLSMV), mean and variance adjusted unweighted least squares (ULSMV) and Bayesian estimation methods were compared in terms of outcome variables. The number of iterations in ML, ULSMV and WLSMV methods is limited to 1,000, which is the default value of Mplus.

When the Bayesian estimation method was used, informative and non-informative priors could be used. Informative priors can be used if the researcher has information about the distribution of parameters. However, non-informative priors can be used if the researcher does not have information about the distribution of parameters (B. O. Muthén & Asparouhov, 2012). Non-informative priors, which is the default in Mplus used in this study, were determined as follows: for indicators, (τ) ~N (0, ∞); for factor loadings (λ) ~N (0,5); for regression coefficients (β) ~N (0,5); for latent response variable's mean / intersection (α) ~N (0, ∞), and, for latent response variable's variance, ~inverse Gamma (-1,0) was used (L. K. Muthén & Muthén, 2015). In this study, tetrachoric correlation matrix was used to conduct CFA because of the binary data.

2.5. Outcome Variables

Firstly, non-convergence solutions were investigated. Following this, improper solutions were examined. If the factor loadings of the indicators are -1.00 and smaller or +1.00 and greater, then this solution was treated as an improper solution and excluded from further analysis.

In order to compare data obtained from the simulation study, relative percentage bias (RPB) values were used (DiStefano & Morgan, 2014; Flora & Curran, 2004; Jin et al., 2016; Lei, 2009; Liang & Yang, 2014). The equation for RPB can be formulated as below:

$$RPB = \frac{\hat{\theta}-\theta_{True}}{\theta_{True}}.100\% \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

Here, $\widehat{\boldsymbol{\theta}}$ is the mean of sample estimates over 1,000 replications, whereas $\boldsymbol{\theta_{True}}$ presents the true value. When the formula of RPB value is examined, it is seen that the value calculated is a percentage. In studies where RPB values are used, absolute values of RPB greater than 10 are taken as the evaluation criteria (Curran et al., 1996; Flora & Curran, 2004; Rhemtulla et al., 2012). Similarly, in this research, RPB values greater than 10 were labeled biased. In this study, RPB for interfactor correlation was calculated only for φ = 0.3 and φ = 0.6. Because of the zero divided problem ($\boldsymbol{\theta_{True}}$=0 in Equation 1), RPB was calculated only for φ = 0.3 and φ = 0.6 conditions.

When reporting RPB, averages of the indicators were calculated. The mean RPB value of the indicators is demonstrated in graphs. In addition, RPB values for each item are given in a table in Appendix 3.

Coverage rate also was used to compare estimation methods' performances. Coverage rate examines the inclusion of the real parameter value of the confidence interval to be established around the parameter estimation. For this purpose, a 95% confidence interval was created for each estimation using the standard error of the estimation, and whether the real parameter value was in this interval was examined. Collins et al. (2001) suggest a coverage rate less than 90% is problematic. Therefore, in the present study, the cut-off point of the coverage rate was 90%.

For the relative bias of standard errors, the relative standard error bias (r-seb) was also calculated. For this:

$$r - seb = \frac{\frac{1}{n_{rep}}\sum_{t=1}^{n_{rep}} \widehat{se}(\hat{\theta}_{pt})}{sd(\hat{\theta}_{pt})} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2$$

equality was used. Where $\widehat{se}(\hat{\theta}_{pt})$, standard error of parameter p for t. replication, $sd(\hat{\theta}_{pt})$ is standard deviation of parameter p for t. replication. R-seb value was classified by Holtmann et al. (2016). Holtmann et al. (2016) as 5/6 < r-seb <6/5 negligible, 2/3 < r-seb <5/6 ve 6/5 < r-seb <3/2 medium, and r-seb <2/3 or r-seb >3/2, large. In the present study, r-seb values which were negligible, and medium were considered as acceptable.

3. RESULT

In this section, results of the simulation study are provided according to the outcome variables.

3.1. Non-convergence and Improper Solutions

Non-convergence was encountered in 15 datasets (0.01%) of 144,000 datasets mostly in 200 sample sizes for the ML method. One of these datasets is in the 500 sample size and the other 14 are in the 200 sample size. The ML method was not converged under conditions where average factor loading is 0.4, indicators follow skewed distribution (right or left) and the number of indicators is 10.

Non-convergence was encountered in 19 datasets (0.01%) of 144,000 datasets, mostly in 200 sample sizes for the Bayesian method. The Bayesian method was not converged under conditions where the number of factors was two, average factor loading was 0.4 and the number of indicators was 10.

The ULSMV and WLSMV methods have more non-convergent datasets than ML and Bayesian methods. Non-convergence was encountered in 2,755 datasets (1.91%) of 144,000 datasets for the ULSMV method. When the properties of the non-converged datasets were examined, it was observed that it occurs mostly where the sample size is 200 and indicators are skewed in the ULSMV method. All non-converged datasets occurred under conditions where the average factor loading was 0.4.

The WLSMV method has non-convergent solutions in 2,856 datasets (1.98%). It was observed that non-convergence occurred mostly in conditions where the sample size is 200 and average factor loading is 0.4. Non-convergent solutions are detailed in Appendix 1.

The Bayesian method has no improper solution. ML has improper solutions in 162 datasets under conditions where sample size is 200, average factor loading is 0.4, the number of indicators is 10, which were skewed. ML has improper solutions for only two factors conditions.

In the ULSMV method, there were improper solutions in 1,534 datasets (1.06%). It was observed that these datasets generally emerged under conditions where the number of indicators is 10, sample size is 200, average factor loading is 0.4, and indicators were skewed. In the WLSMV method, there were improper solutions in 1,877 datasets (1.30%). It was observed that these datasets generally emerged under conditions where the number of indicators is 10, sample size is 200, average factor loading is 0.4, and indicators were skewed. The number of datasets with the improper solution is detailed in Appendix 2.

3.2. Relative Percentage Bias

3.2.1. *Relative percentage bias of factor loadings*

RPB values obtained from simulation conditions are presented in Figure 2. In addition, the maximum, average and minimum values of the RPB obtained from the items are given in Appendix 3 for researchers who want to examine them.

When the RPB values obtained from the estimation methods for factor loadings (Figure 2) are examined, it can be said that all methods have an acceptable bias for sample sizes of 500 and 1,000. It was observed that RPB values of all estimation methods are less than 10 under conditions with average factor loading of 0.7 in a sample size of 200. However, RPB can be smaller than -10 where average factor loading is 0.4 in the sample size of 200. Considering the conditions with an average factor loading of 0.4 and the number of indicators of 10, the RPB of the Bayesian method is less than -10 where the indicators are skewed and factorial structure consist of two factors (for $\varphi = 0$ and 0.3). For the same conditions, except for the number of items, the RPB of the WLSMV is less than -10 where the number of indicators is 20 and two factorial structure (for φ = 0 and 0.3). Increasing the number of indicators in skewed distributions reduced bias. In addition, increasing the interfactor correlation reduced bias. ML and ULSMV methods have acceptable bias in all conditions. The WLSMV method's RPB is less than -10 in just one condition (mean factor loading is 0.4, right-skewed indicators, the number of indicators is 20, and sample size is 200). All of the methods have negative bias.

3.2.2. *Relative percentage bias of interfactor correlations*

The results obtained from simulation conditions for interfactor correlations are presented in Figure 3. In addition, for researchers who want to examine further detail, values are given in table in Appendix 4. When the RPB values of the methods for interfactor correlation (Figure 3) are examined, RPB, obtained from all methods is within acceptable limits under conditions with an average factor loading of 0.7. However, as the sample size decreases under the conditions with an average factor loading of 0.4, the RPB of the φ parameter obtained from the methods may go beyond the limits. The RPB obtained from the Bayesian method was estimated to be less than the required value for both models under conditions where the sample size was 200 and 500, and the average factor loading was 0.4. RPB values of ML, ULSMV and WLSMV methods are within acceptable limits under conditions where sample size is 500 and average factor loading is 0.4. Under the conditions where average factor load was 0.4, sample size was 200, and skewed distribution, the number of items was increased, the RPB values of ML, ULSMV and WLSMV methods increased to acceptable range. Under the conditions where average factor loading was 0.4, sample size was 200, and normal distribution, the RPB values of ML, ULSMV and WLSMV methods were within acceptable limits.

3.3. Coverage Rate

3.3.1. *Coverage rate of factor loadings*

The coverage rates obtained from the simulation conditions are presented in Figure 4. In addition, the maximum, minimum and average values of the coverage rates obtained from the items are given in Appendix 5 for researchers who want to study the detail. When the coverage rates of the methods are examined according to the simulation conditions, it was observed that the coverage rates of the estimation methods decreased under conditions where the sample size is 200 and the items were skewed. When average factor loading increases to 0.7, the coverage rates of estimation methods increase for a sample size of 200. The coverage rate of ULSMV and WLSMV is below 90% under the conditions where sample size is 200, average factor loading is 0.4 and the items are skewed. It can be said that the Bayesian method performs better than the others under these conditions. Increasing the number of items increased the coverage rate of ML. The coverage rate of the Bayesian method is less than 90% for the conditions where the factor structure is unidimensional, the number of items is 20 and average factor loading is 0.4.

Under the conditions where the sample size is 500, the coverage rate of all methods, except for Bayesian, is over 90% for all models. However, the Bayesian method can fall below 90% in unidimensional structures. With the increase in the average factor loading, the performance of the Bayesian method in unidimensional structures increases. When the simulation conditions where the sample size is 1,000 are examined, the Bayesian method has a lower coverage rate in unidimensional structures than the other structures under conditions where average factor loading is 0.4. The coverage rates of the ML, ULSMV and WLSMV methods are adequate for all models for the conditions where the indicators are normal and distribution skewed. Under conditions with an average factor loading of 0.7, the coverage rates of all other methods, except the ML method, are sufficient. The coverage rate of ML is below 90% for some models under conditions where the number of indicators is 10 or 20. Interestingly, the increase in sample size reduced the coverage rate of ML for these conditions.

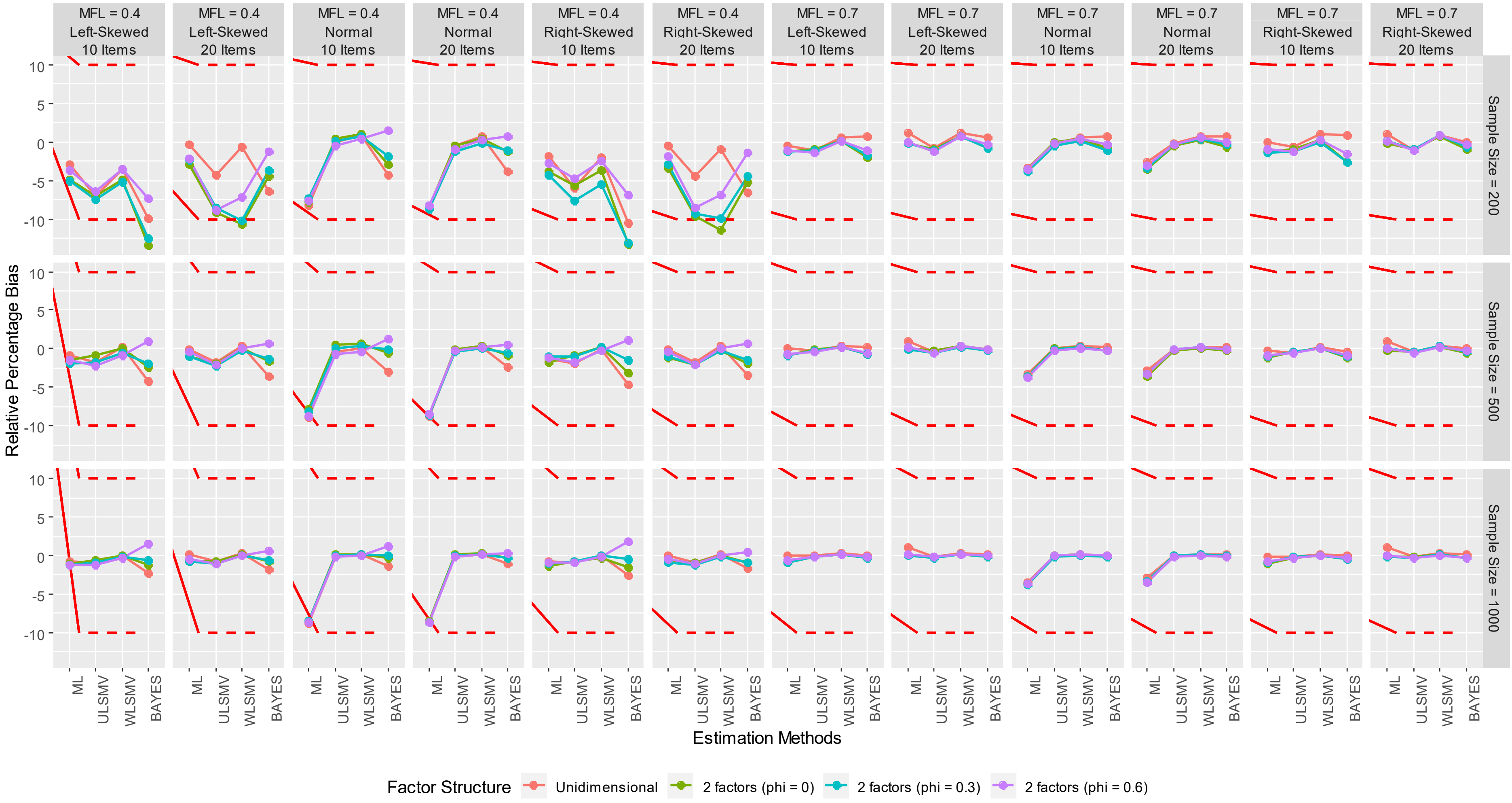


Figure 2. *Relative percentage bias (RPB) of factor loadings*

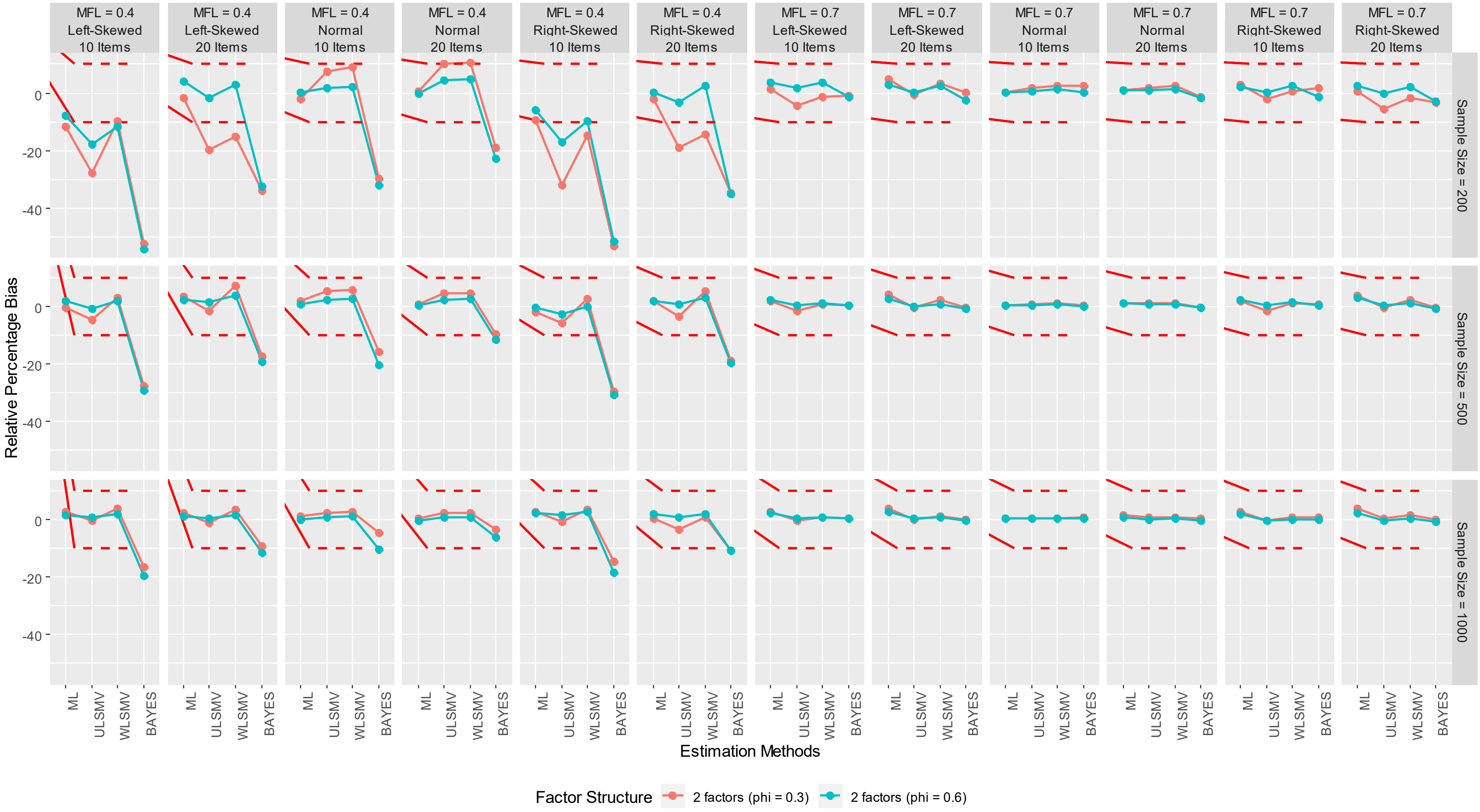


Figure 3. *Relative percentage bias (RPB) of interfactor correlations*

3.3.2. *Coverage rate of interfactor correlations*

The coverage rate of the methods for interfactor correlation are presented in Figure 5, and the numerical values are presented in a table in Appendix 6. Coverage rates obtained from all estimation methods are 90% and above, under conditions where the sample size is 1,000 and average factor loading is 0.7. Under the conditions where the sample size is 200 and average factor loading is 0.7, the coverage rates of the methods are above 90% under the conditions where indicators follow normal distribution. However, the performance of the WLSMV method decreased as the φ parameter decreases under conditions where the indicators distributions are skewed. ML and Bayesian methods have a coverage rate of over 90% under conditions where the sample size is 200, indicators are skewed and average factor loading is 0.7. The ULSMV method had a coverage rate of over 90% with the increase of the φ parameter under conditions where the sample size is 200, indicators were skewed and average factor loading is 0.7, and remained below 90% under conditions where φ was 0.

The conditions where indicators follow normal distribution and average factor loading is 0.4, increasing the sample size increased coverage rate of ML, ULSMV and WLSMV. In addition, the coverage rate of the Bayesian method is less than 90% where the φ parameter is 0.6. Decreasing the φ parameter increased the coverage rate of the Bayesian method. The conditions where indicators are normally distributed, the average factor loading is 0.4 and sample size is 200, increasing the interfactor correlation (φ) increased the coverage rate of ULSMV and WLSMV. With the increase in the number of indicators in these conditions, the coverage rate of ML increased but it was not affected by the magnitude of the interfactor correlation.

Under conditions where the average factor loading is 0.4, indicators follow skewed distribution, sample size is 200 and the number of items is 10, the coverage rate of the Bayesian method alone (for φ = 0 and 0.3) is higher than 90%, while when the number of indicators increased to 20, the coverage rate of ML is about 90%. Under all conditions where the average factor loading is 0.4, items have skewed distribution and sample size is 500, the coverage rate of the ULSMV and WLSMV methods is higher than 90% only if the φ parameter is 0.6. For these conditions, the coverage rate of ML is higher than 90%. With the increase in the number of items, the coverage rate of ML also increased. While the Bayesian method has a coverage rate of less than 90% under conditions where the φ parameter is 0.6, under conditions where the φ parameter is 0 or 0.3, the coverage rate of the Bayesian method is higher than 90%.

3.4. Relative Standard Error Bias

3.4.1. *Relative standard error of factor loadings*

The r-seb values obtained from the simulation conditions are presented in Figure 6. In addition, maximum, minimum values and averages of the r-seb values obtained from the indicators are given in Appendix 7. In all conditions where the sample size is 500 and 1,000, all estimation methods have an acceptable r-seb value for all models. However, the WLSMV method has a large r-seb value in all two-dimensional models, except for unidimensional structures under 20-indicators conditions with a skewed distribution with an average factor loading of 0.4 in 200 sample size.

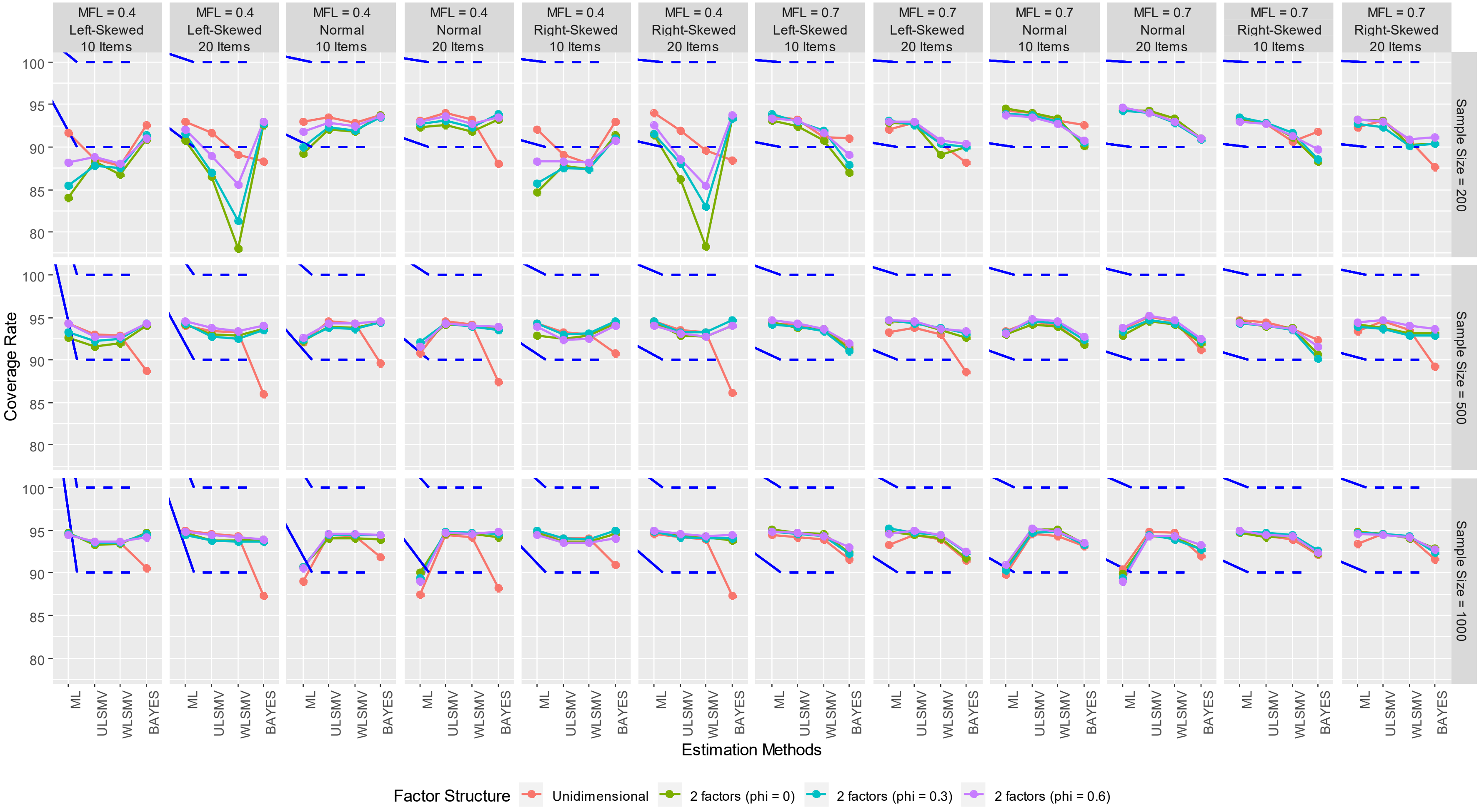


Figure 4. *Coverage rate of factor loadings*

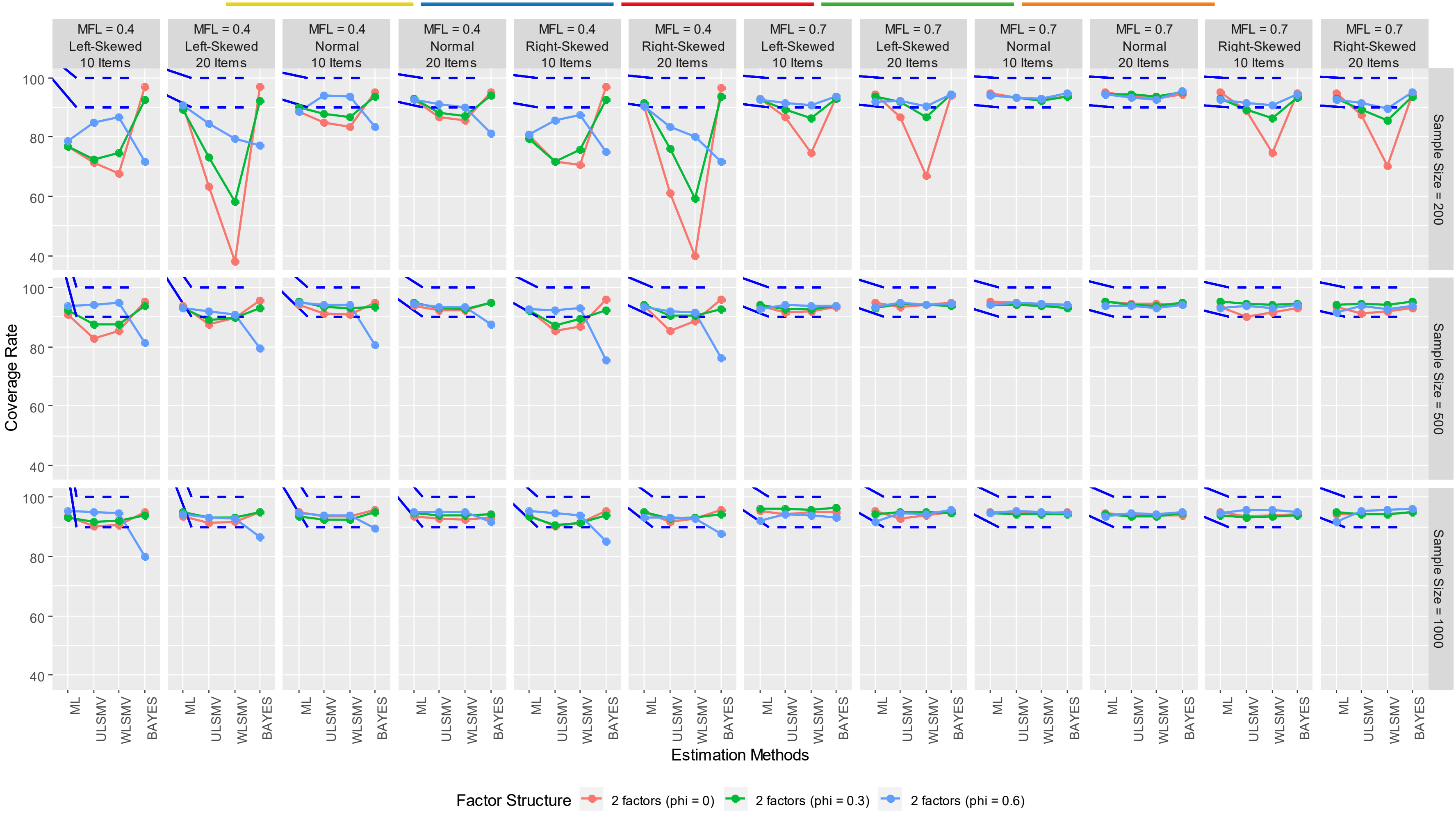


Figure 5. *Coverage rate of interfactor correlations*

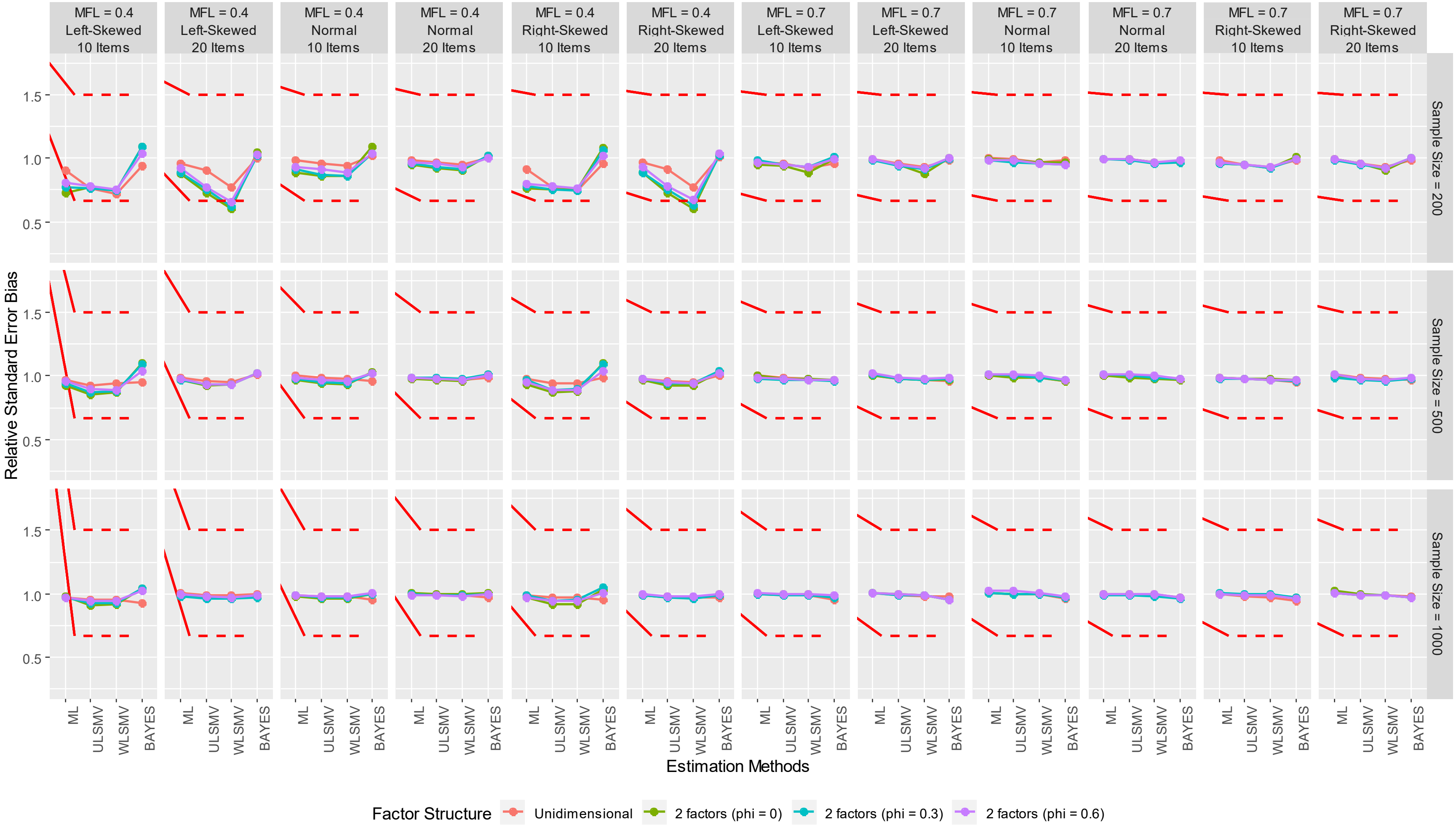


Figure 6. Relative standard error bias (r-seb) of factor loadings

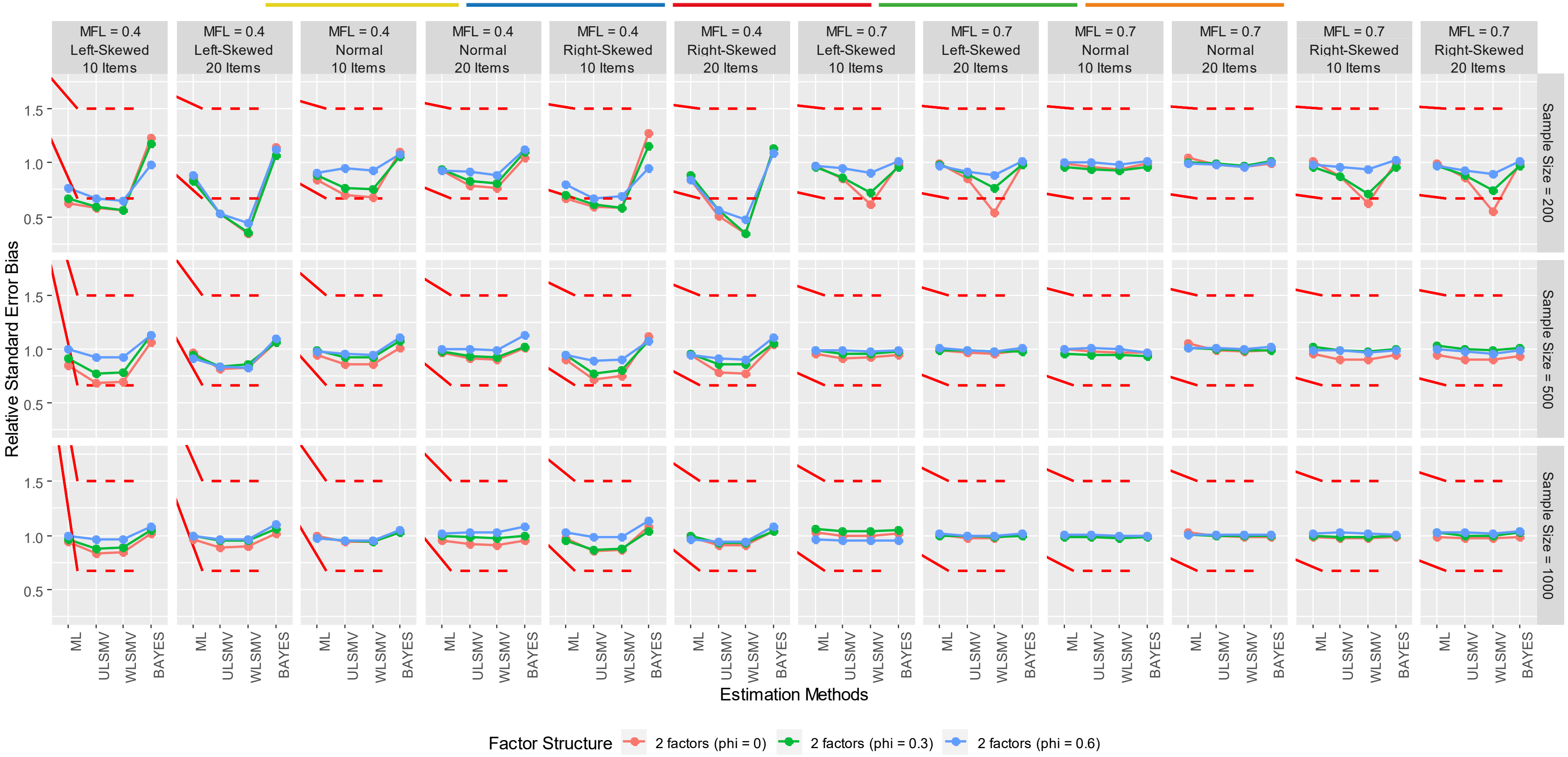


Figure 7. Relative standard error bias (r-seb) of interfactor correlations

3.4.2. *Relative standard error of interfactor correlations*

The r-seb values of the methods for interfactor correlation are presented in Figure 7 and the numerical values are given in the table in Appendix 8. When Figure 7 is examined, it can be said that the r-seb values of all estimation methods are acceptable under all conditions where sample sizes are 500 and 1,000. However, the r-seb values of the estimation methods are out of range for decreasing the sample size and the indicators became skewed.

The r-seb values of the estimation methods are acceptable under conditions where the sample size is 200 and the average factor loading is 0.7. In cases where the items are skewed, the r-seb values of the WLSMV method are out of the acceptable range when φ parameter is 0. The r-seb values of the other methods are acceptable under these conditions.

The r-seb values of the estimation methods are acceptable under conditions where the sample size is 200, the average factor loading is 0.4 and indicators follow normal distribution. In these conditions, except for distribution of indicators, the r-seb values of the Bayesian method are acceptable under conditions where indicators follow skewed distribution. Increasing the number of items under these conditions, the r-seb values of the ML method increased to an acceptable range. The r-seb values of the ULSMV and WLSMV methods are unacceptable for these conditions. In these conditions, ML and Bayesian methods perform better in terms of r-seb values.

4. DISCUSSION

In this study, CFA estimation methods were compared by manipulating sample size, distribution of data, test length, average factor loading, and factor structure for binary data.

4.1. Non-convergence and Improper Solutions

Non-convergence frequently encounters datasets which have a two-factor structure and consist of skewed indicators for the ULSMV and WLSMV estimation methods. These methods have a less converged problem in unidimensional structures than in two-factor structures (even if items are skewed). The increase in the number of items for conditions where items are skewed decreases the non-convergence datasets for ULSMV and WLSMV. It can be said that all estimation methods converge when the average factor loading is 0.7. In other words, the ULSMV and WLSMV estimation methods are mostly non-convergent for small sample size, low average factor loading, short test length, two-dimensional models, and the magnitude of interfactor correlation is small. This result is consistent with the study by Moshagen and Musch (2014). In a study comparing MLR and WLSMV estimation methods conducted by Li (2016a), consisting of 4, 6, 8, and 10 categories indicators with skewness coefficients ranging from 1.01-1.31, it was reported that WLSMV converged under all conditions and there were no improper solutions. It is thought that the differentiation of the number of categories of indicators and the skewness of the indicators may have caused differentiation between the results. Nestler (2013) states that the DWLS (WLSMV) method had 3.7% non-convergence in sample size of 250. The present study is similar to the study conducted by Flora and Curran (2004) and Nestler (2013) in terms of non-convergence.

There are no improper solutions in the Bayesian method, whereas there are a few in the ML method. Liang and Yang (2014) also reported that there were no improper solutions in binary data in a simulation study using non-informative priors. The result obtained in this respect is consistent with the Liang and Yang's (2014) study. The ULSMV and WLSMV methods have more improper solutions under conditions where average factor loading is 0.4 and sample size is 200. There was a decrease in the number of improper solutions under conditions where the average factor load was 0.4, sample size was 200 and the indicators followed a normal distribution. The number of improper solutions is very close to 0 even if the sample size is small under conditions where the average factor load was 0.7.

When non-convergence and improper solution results are evaluated, it can be said that ML and Bayesian methods perform better than ULSMV and WLSMV methods. It was observed that the number of non-convergence and improper solutions of the methods increased where the average factor loading was low, test length was short and the distribution of indicators were skewed, the factor structure was not unidimensional, and the sample size was small.

4.2. Relative Percentage Bias of Factor Loadings

When the RPB values calculated via factor loadings of the estimation methods were examined, the Bayesian and WLSMV methods may give biased results under conditions where the sample size is 200, average factor loading is 0.4, and items are skewed. The conditions where the sample size is 200, average factor loading is 0.4, and items are skewed show that increasing the number of indicators and number of factors and decreasing the interfactor correlation decreased the WLSMV estimation method's RPB, and decreasing the number of indicators and the interfactor correlation while increasing the number of factors decreased the RPB value of the Bayesian method. This result is similar to the findings obtained by Nalbantoğlu Yılmaz (2019) in continuous data. She stated that the WLS method has larger RPB values for small samples. In addition, this result is consistent with the research conducted by Moshagen and Musch (2014) and Lei (2009). Moshagen and Musch (2014) report that RPB was less than 10% for unidimensional structures, while Lei (2009) states that RPB values of the ML and WLSMV methods were less than 10%. Flora and Curran (2004) report that the RPB values of robust WLS estimation methods did not exceed 10%. But in the present study, the RPB value of WLSMV is more than 10% in conditions where the sample size is 200, the average factor loading is 0.4, the distribution of 20 indicators are left-skewed, and the interfactor correlation is 0 and 0.3. This difference may be due to the fact that the factor loadings are not equal for all indicators in the present study. In addition, Flora and Curran (2004) make y* (latent continuous variable) skewed. However, in the present study, the latent variable (y*) was generated to follow a normal distribution. Indicators were skewed and analyses were performed. It is thought that the differentiation of the results may have been due to this differentiation.

4.3. Relative Percentage Bias of Interfactor Correlations

When the RPB values of the φ parameter, which is interfactor correlation, are examined, the RPB performance of all methods is sufficient in all conditions with average factor loading of 0.7. However, the RPB performance of the ML, ULSMV and WLSMV methods decreased when the sample size decreased and the skewness of the items increased under conditions where average factor loading is 0.4. Beauducel and Herzberg (2006) state that interfactor correlation size is more effective on the performance of the estimation methods. In this respect, it can be said that the present study is similar to Beauducel and Herzberg's (2006) study. The RPB value of the Bayesian method is higher than 10% in almost all conditions where the average factor loading is 0.4, sample size is 1,000, distribution of indicators normal and the number of indicators is 20. This result is consistent with the findings of Liang and Yang (2014).

4.4. Coverage Rate of Factor Loadings

When the coverage rates calculated via factor loadings of the estimation methods were examined, it was observed that the increase in sample size and average factor loading increased the performance of the estimation methods. It can be said that the coverage rate of the ULSMV and WLSMV methods are not sufficient in the conditions where the distribution of indicators is skewed, sample size is small, and average factor loadings is low. This result is consistent with the findings of the simulation study conducted by Forero et al. (2009). In addition, Koğar and Yılmaz Koğar (2015) stated that ULS and DWLS methods have less standard errors when compared to the ML method. The difference may have originated from variables that were not examined in current study included in the real data set. In the simulation study conducted by Wolf et al. (2013), it was reported that ML had sufficient coverage rates under all conditions studied. However, in this study, the data were generated as normal and continuous. It can be said that there may be a difference in this respect with the results of the present study. The coverage rate of the Bayesian method is less than 90% for the conditions where the model is unidimensional and average factor loading is 0.4. Önen (2019) states that the coverage rate of the Bayesian method is sufficient for all simulation conditions. In the present study, the difference may have arisen since non-informative priors for Bayesian estimations.

4.5. Coverage Rate of Interfactor Correlations

Coverage rates calculated for interfactor correlations for the WLSMV method remained below 90% under conditions where the correlation between the dimensions was 0 and a small sample size. The performance of the WLSMV method decreased as the φ parameter decreased under conditions where the indicators followed a skewed distribution. The ML and Bayesian methods had a coverage rate of over 90% under conditions where sample size is 200, average factor load is 0.7 and items are skewed. Li (2016b) likewise reported that the coverage rate of the MLR method is adequate, and that the WLSMV method may have a coverage rate of less than 90% in skewed distributions. It can be said that the current research findings are consistent with this study. The ULSMV method's coverage rate remained below 90% as interfactor correlation decreases under conditions of sample size of 200 and skewed distribution of items. Under conditions where the average factor loading is 0.4, sample is small and the indicators follow normal distribution, the coverage of the ULSMV and WLSMV methods decreases as interfactor correlation decreases. In the case of indicators with average factor loading of 0.4, the performances of other methods remained below 90%, except for the Bayesian method. However, with the increase in the sample size, the coverage rate of the methods increased.

4.6. Relative Standard Error Bias of Factor Loadings

When r-seb values are examined for factor loadings, it can be said that the methods perform sufficiently in most of the conditions. The WLSMV method went beyond the acceptable limits for r-seb values under conditions where the average factor loading is 0.4, the sample size is 200 and the number of indicators are 20, and is within the acceptable range under other conditions. Other methods are acceptable in all conditions. In the simulation study performed by Xu (2019), the MLR method has sufficient relative bias in normal, mild non-normal and moderate non-normal data. However, the Bayesian method has a relative bias greater than 10% under moderate non-normal conditions. In the present study, the Bayesian method has sufficient r-seb value under all conditions. In the study conducted by Xu (2019), the data was produced as a correlation matrix. In addition, the factor loadings of the indicators were fixed to 0.7. It can be said that the differentiation may have originated from here.

4.7. Relative Standard Error Bias of Interfactor Correlations

When the r-seb values are examined for interfactor correlations, all methods are in the acceptable range in 500 and 1,000 sample sizes, while the WLSMV method in the sample size of 200 may be out of the acceptable range in skewed distributions.

When the results of the research are evaluated in general, it can be said that the increase in the average factor loading and the sample size have a positive effect on the performances of the estimation methods. The increase in the number of indicators did not cause much difference for the indicators which follow normal distribution, but it affected the estimations of the methods for the indicators which followed skewed distribution. According to the research findings, it can be said that the methods are sufficient to estimate the average factor loading and the interfactor correlations, regardless of the estimation method used in most of the conditions where the average factor loading is 0.7. However, as the average factor loading was 0.4, the number of skewed indicators increased, the sample size decreased and the interfactor correlations decreased, the performance of the methods decreased. Especially in small samples, the interfactor correlation was lower in the case of skewed indicators than indicators which follow normal distribution.

According to the research findings, it can be said that any estimation method can be chosen under conditions where sample size is 500 or 1000 and average factor loading is 0.7. The performance of the estimation methods differs in conditions with a sample size of 200. Therefore, the conditions where sample size is 200, the average factor loading is 0.4, indicators follow normal distribution, and the structure is unidimensional, it is recommended to use ML, ULSMV or WLSMV.

However, if the indicators are skewed, it can be recommended that to use ML or ULSMV estimation method. As the interfactor correlation decreases, the performance of the estimation methods to estimate the interfactor correlations decreases in small samples. Therefore, expanding the sample can be considered in such a case. According to the research findings, there is no method that makes the most accurate estimation under all conditions. However, it can be suggested that to use the ULSMV estimation method because it is observed that it has sufficient performance under more conditions.

Declaration of Conflicting Interests and Ethics

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

ORCID

Abdullah Faruk KILIÇ https://orcid.org/0000-0003-3129-1763
İbrahim UYSAL https://orcid.org/0000-0002-6767-0362
Burcu ATAR https://orcid.org/0000-0003-3527-686X

5. REFERENCES

Acar-Güvendir, M., & Özer-Özkan, Y. (2015). Türkiye'deki eğitim alanında yayımlanan bilimsel dergilerde ölçek geliştirme ve uyarlama konulu makalelerin incelenmesi [The examination of scale development and scale adaptation articles published in Turkish academic journals on education]. *Electronic Journal of Social Sciences*, *14*(52), 23–33. https://doi.org/10.17755/esosder.54872

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anıl, D., Güzeller, C. O., Çokluk, Ö., & Şekercioğlu, G. (2010). Level determination exam (SBS-2008) the determination of the validity and reliability of 7th grade mathematics sub-test. *Procedia-Social and Behavioral Sciences*, *2*(2), 5292-5298. https://doi.org/10.1016/j.sbspro.2010.03.863

Babakus, E., Ferguson, C. E., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, *24*(2), 222-228. https://doi.org/10.2307/3151512

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(2), 186-203. https://doi.org/10.1207/s15328007sem1302_2

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons, Inc.

https://doi.org/10.1002/9781118619179

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. *Psychometrika*, *50*(2), 229-242. https://doi.org/10.1007/BF02294248

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). Routledge.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Lawrence Erlbaum Associates.

Crocker, L., & Algina, J. (2008). *Introduction of classical and modern test theory*. Cengage Learning.

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

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 425-438. https://doi.org/10.1080/10705511.2014.915373

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*(2), 309-326. https://doi.org/10.1111/j.2044-8317.1994.tb01039.x

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286–299. https://doi.org/10.1037/1040-3590.7.3.286

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A monte carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 625-641. https://doi.org/10.1080/10705510903203573

Gilbert, N. (1999). Simulation: A new way of doing social science. *American Behavioral Scientist*, *42*(10), 1485–1487. https://doi.org/10.1177/00027642990042010002

Gorsuch, R. L. (1974). *Factor analysis*. W. B. Saunders.

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 108-120. https://doi.org/10.1080/10705519709540064

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265-275. https://doi.org/10.1037/0033-2909.103.2.265

Hallquist, M., & Wiley, J. (2017). *MplusAutomation: Automating Mplus Model Estimation and Interpretation* [Computer software]. https://cran.r-project.org/web/packages/MplusAutomation/

Harrison, R. L. (2010). Introduction to Monte Carlo Simulation. *AIP Conference Proceedings,* *1204*, 17–21. https://doi.org/10.1063/1.3295638

Holtmann, J., Koch, T., Lochner, K., & Eid, M. (2016). A Comparison of ML, WLSMV, and Bayesian methods for multilevel structural equation models in small samples: A simulation study. *Multivariate Behavioral Research*, *51*(5), 661-680. https://doi.org/10.1080/00273171.2016.1208074

Jin, S., Luo, H., & Yang-Wallentin, F. (2016). A simulation study of polychoric instrumental variable estimation in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 680-694. https://doi.org/10.1080/10705511.2016.1189334

Kılıç, A. F., & Kelecioğlu, H. (2016). TEOG ortak ve mazeret sınavındaki Türkçe ve matematik alt testlerinin psikometrik özelliklerinin karşılaştırılması [The comparison of psychometric properties of standardised and make up maths and Turkish subtest questions in TEOG]. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(1), 33–58. https://doi.org/10.21031/epod.14532

Kline, R. B. (2016). *Principle and practice of structural equation modeling* (4th ed.). The Guilford.

Koğar, H., & Yılmaz Koğar, E. (2015). Comparison of different estimation methods for categorical and ordinal data in confirmatory factor analysis. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2), 351-364. https://doi.org/10.21031/epod.94857

Lei, P. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity*, *43*(3), 495–507. https://doi.org/10.1007/s11135-007-9133-z

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

Li, C.-H. (2016b). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods, 21*(3), 369–387. https://doi.org/10.1037/met0000093

Liang, X., & Yang, Y. (2014). An evaluation of WLSMV and Bayesian methods for confirmatory factor analysis with categorical indicators. *International Journal of Quantitative Research in Education*, *2*(1), 17-38. https://doi.org/10.1504/IJQRE.2014.060972

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology*, *13*(3), 1-50. http://www.jattjournal.com/index.php/atp/article/view/48366/39234

Morata-Ramirez, M. de los A., & Holgado-Tello, F. P. (2013). Construct validity of likert scales through confirmatory factor analysis: A simulation study comparing different methods of estimation based on Pearson and polychoric correlations. *International Journal of Social Science Studies*, *1*(1), 54-61. https://doi.org/10.11114/ijsss.v1i1.27

Moshagen, M., & Musch, J. (2014). Sample size requirements of the robust weighted least squares estimator. *Methodology*, *10*(2), 60-70. https://doi.org/10.1027/1614-2241/a000068

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Chapman & Hall.

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. https://www.statmodel.com/download/webnotes/CatMGLong.pdf

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171–189. https://doi.org/10.1111/j.2044-8317.1985.tb00832.x

Muthén, L. K., & Muthén, B. O. (2012). *Mplus statistical modeling software: Release 7.0* [Computer software]. Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide* (7th ed.). Muthén & Muthén.

Nalbantoğlu Yılmaz, F. (2019). Comparison of different estimation methods used in confirmatory factor analyses in non-normal data: A monte carlo study. *International Online Journal of Educational Sciences*, *11*(4), 131-140. https://doi.org/10.15345/iojes.2019.04.010

Nestler, S. (2013). A monte carlo study comparing PIV, ULS and DWLS in the estimation of dichotomous confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 127-143. https://doi.org/10.1111/j.2044-8317.2012.02044.x

Önen, E. (2019). A comparison of frequentist and Bayesian approaches: The power to detect model misspecifications in confirmatory factor analytic models. *Universal Journal of Educational Research*, *7*(2), 494–514. https://doi.org/10.13189/ujer.2019.070223

Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology*, *46*(2), 273–286. https://doi.org/10.1111/j.2044-8317.1993.tb01016.x

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Lawrence Erlbaum Associates.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Şahin, M. G., & Boztunç Öztürk, N. (2018). Eğitim alanında ölçek geliştirme süreci: Bir içerik analizi çalışması [Scale development process in educational field: A content analysis research]. *Kastamonu Education Journal*, *26*(1), 191-199. https://doi.org/10.24106/kefdergi.375863

Shi, D., DiStefano, C., McDaniel, H. L., & Jiang, Z. (2018). Examining chi-square test statistics under conditions of large model size and ordinal Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(6), 924-945. https://doi.org/10.1080/10705511.2018.1449653

Stevens, J. P. (2009). *Applied multivariate statistics for the social science* (5th ed.). Taylor & Francis.

Streiner, D. L. (1994). Figuring out factors: The use and misuse of factor analysis. *Canadian Journal of Psychiatry*, *39*(3), 135-140. https://doi.org/10.1177%2F070674379403900303

Tabachnik, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Pearson.

Thissen, D., Wainer, H., & Wang, X.-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, *31*(2), 113–123. https://doi.org/10.1111/j.1745-3984.1994.tb00437.x

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety.

National Institutes of Health, 76(6), 913-934. https://doi.org/10.1177/0013164413495237

Xu, M. (2019). *A comparison of frequentist and Bayesian approaches for confirmatory factor analysis* (Publication No. 27534819) [Doctoral dissertation, The Ohio State University]. ProQuest Dissertations and Theses Global.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(3), 392–423. https://doi.org/10.1080/10705511.2010.489003

6. APPENDIX

Appendix 1. Number of datasets having convergence failure

Sample Size	Model	Method	Mean Factor Loading = 0.4						Mean Factor Loading = 0.7					
			Left-Skewed		Normal		Right-Skewed		Left-Skewed		Normal		Right-Skewed	
			10	20	10	20	10	20	10	20	10	20	10	20
200	Unidimensional	ML	-	-	-	-	-	-	-	-	-	-	-	-
200		ULSMV	41	32	2	3	50	26	-	-	-	-	-	-
200		WLSMV	29	41	2	3	49	30	1	-	-	-	-	-
200		BAYES	1	-	-	-	2	-	-	-	-	-	-	-
200	2 factors (φ = 0)	ML	2	-	-	-	2	-	-	-	-	-	-	-
200		ULSMV	299	101	81	11	290	120	-	-	-	-	-	-
200		WLSMV	282	157	85	11	306	158	-	-	-	-	1	-
200		BAYES	3	2	1	-	-	3	-	-	-	-	-	-
200	2 factors (φ = 0.3)	ML	2	-	-	-	4	-	-	-	-	-	-	-
200		ULSMV	256	101	48	6	282	110	-	-	-	-	-	-
200		WLSMV	240	157	49	6	270	160	1	-	-	-	1	-
200		BAYES	-	-	-	-	1	-	-	-	-	-	-	-
200	2 factors (φ = 0.6)	ML	4	-	-	-	-	-	-	-	-	-	-	-
200		ULSMV	211	1-3	27	5	186	87	-	-	-	-	-	-
200		WLSMV	193	1-2	27	5	165	109	-	-	-	-	-	1
200		BAYES	1	2	-	-	-	2	-	-	-	-	-	-
500	Unidimensional	ML	-	-	-	-	-	-	-	-	-	-	-	-
500		ULSMV	4	-	-	-	-	-	-	-	-	-	-	-
500		WLSMV	5	-	-	-	-	-	-	-	-	-	-	-
500		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
500	2 factors (φ = 0)	ML	1	-	-	-	-	-	-	-	-	-	-	-
500		ULSMV	65	5	1	-	70	1	-	-	-	-	-	-
500		WLSMV	54	4	1	-	57	1	-	-	-	-	-	-
500		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
500	2 factors (φ = 0.3)	ML	-	-	-	-	-	-	-	-	-	-	-	-
500		ULSMV	43	6	-	-	34	2	-	-	-	-	-	-
500		WLSMV	30	5	-	-	18	1	-	-	-	-	-	-
500		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
500	2 factors (φ = 0.6)	ML	-	-	-	-	-	-	-	-	-	-	-	-
500		ULSMV	17	1	-	-	16	5	-	-	-	-	-	-
500		WLSMV	13	1	-	-	14	4	-	-	-	-	-	-
500		BAYES	-	-	-	-	1	-	-	-	-	-	-	-
1000	Unidimensional	ML	-	-	-	-	-	-	-	-	-	-	-	-
1000		ULSMV	-	-	-	-	-	-	-	-	-	-	-	-
1000		WLSMV	-	-	-	-	-	-	-	-	-	-	-	-
1000		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
1000	2 factors (φ = 0)	ML	-	-	-	-	-	-	-	-	-	-	-	-
1000		ULSMV	3	-	-	-	2	-	-	-	-	-	-	-
1000		WLSMV	3	-	-	-	2	-	-	-	-	-	-	-
1000		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
1000	2 factors (φ = 0.3)	ML	-	-	-	-	-	-	-	-	-	-	-	-
1000		ULSMV	2	-	-	-	-	-	-	-	-	-	-	-
1000		WLSMV	2	-	-	-	-	-	-	-	-	-	-	-
1000		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
1000	2 factors (φ = 0.6)	ML	-	-	-	-	-	-	-	-	-	-	-	-
1000		ULSMV	-	-	-	-	-	-	-	-	-	-	-	-
1000		WLSMV	-	-	-	-	-	-	-	-	-	-	-	-
1000		BAYES	0	0	0	0	0	0	0	0	0	0	0	0

Appendix 2. Number of datasets having inadmissible solution

Sample Size	Model	Method	Mean Factor Loading = 0.4						Mean Factor Loading = 0.7					
			Left-Skewed		Normal		Right-Skewed		Left-Skewed		Normal		Right-Skewed	
			10	20	10	20	10	20	10	20	10	20	10	20
200	Unidimensional	ML	-	-	-	-	-	-	-	-	-	-	-	-
200		ULSMV	9	-	-	-	11	-	-	-	-	-	-	-
200		WLSMV	8	-	-	-	11	-	-	-	-	-	-	-
200		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
200	2 factors (φ = 0)	ML	37	1	-	-	28	-	2	-	-	-	-	-
200		ULSMV	219	24	57	-	205	30	21	-	-	-	26	1
200		WLSMV	239	96	55	-	199	111	53	8	-	-	80	15
200		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
200	2 factors (φ = 0.3)	ML	25	-	1	-	27	1	1	-	-	-	1	-
200		ULSMV	179	20	35	1	185	34	19	1	-	-	25	1
200		WLSMV	169	83	34	1	176	83	17	-	-	-	27	-
200		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
200	2 factors (φ = 0.6)	ML	19	1	-	-	10	-	-	-	-	-	1	-
200		ULSMV	118	5	4	-	129	4	6	3	-	-	12	2
200		WLSMV	88	29	4	-	105	23	3	3	-	-	10	2
200		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
500	Unidimensional	ML	-	-	-	-	-	-	-	-	-	-	-	-
500		ULSMV	-	-	-	-	-	-	-	-	-	-	-	-
500		WLSMV	-	-	-	-	-	-	-	-	-	-	-	-
500		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
500	2 factors (φ = 0)	ML	3	-	-	-	1	-	-	-	-	-	-	-
500		ULSMV	52	-	-	-	35	-	-	-	-	-	-	-
500		WLSMV	50	1	-	-	38	4	-	-	-	-	1	-
500		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
500	2 factors (φ = 0.3)	ML	1	-	-	-	1	-	-	-	-	-	-	-
500		ULSMV	26	-	1	-	22	-	-	-	-	-	1	-
500		WLSMV	22	-	1	-	21	-	-	-	-	-	-	-
500		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
500	2 factors (φ = 0.6)	ML	-	-	-	-	-	-	-	-	-	-	-	-
500		ULSMV	3	-	-	-	5	-	-	-	-	-	-	-
500		WLSMV	-	-	-	-	4	-	-	-	-	-	-	-
500		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
1000	Unidimensional	ML	-	-	-	-	-	-	-	-	-	-	-	-
1000		ULSMV	-	-	-	-	-	-	-	-	-	-	-	-
1000		WLSMV	-	-	-	-	-	-	-	-	-	-	-	-
1000		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
1000	2 factors (φ = 0)	ML	-	-	-	-	1	-	-	-	-	-	-	-
1000		ULSMV	-	-	-	-	2	-	-	-	-	-	-	-
1000		WLSMV	-	-	-	-	2	-	-	-	-	-	-	-
1000		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
1000	2 factors (φ = 0.3)	ML	-	-	-	-	-	-	-	-	-	-	-	-
1000		ULSMV	1	-	-	-	-	-	-	-	-	-	-	-
1000		WLSMV	1	-	-	-	-	-	-	-	-	-	-	-
1000		BAYES	-	-	-	-	-	-	-	-	-	-	-	-
1000	2 factors (φ = 0.6)	ML	-	-	-	-	-	-	-	-	-	-	-	-
1000		ULSMV	-	-	-	-	-	-	-	-	-	-	-	-
1000		WLSMV	-	-	-	-	-	-	-	-	-	-	-	-
1000		BAYES	-	-	-	-	-	-	-	-	-	-	-	-

Appendix 3. Mean, maximum and minimum values of relative percentage bias

Sample Size	Model	MFL	Method	Number of Items = 10									Number of Items = 20								
				Left-Skewed			Normal			Right-Skewed			Left-Skewed			Normal			Right-Skewed		
				Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
200	Unidimensional	0.4	ML	-3.0	0.7	-5.3	-8.2	-5.3	-10.5	-1.8	0.8	-3.9	-0.3	1.4	-2.7	-8.5	-6.7	-10.2	-0.5	1.7	-1.8
200		0.4	ULSMV	-7.0	2.8	-10.8	-0.1	2.6	-2.0	-5.9	1.9	-8.8	-4.3	-2.4	-6.3	-0.5	1.3	-2.3	-4.4	-1.7	-6.8
200		0.4	WLSMV	-3.6	7.8	-6.3	0.9	3.6	-1.1	-2.0	6.7	-5.1	-0.6	3.8	-3.4	0.7	2.6	-0.9	-0.9	5.2	-2.8
200		0.4	BAYES	-10.0	-5.4	-17.1	-4.3	0.3	-14.1	-10.5	-5.3	-14.6	-6.4	-0.7	-28.9	-3.8	0.5	-28.1	-6.6	-1.9	-28.7
200		0.7	ML	-0.4	0.1	-0.8	-3.3	-2.4	-4.0	-0.0	0.5	-0.3	1.2	1.9	0.8	-2.6	-1.8	-3.3	1.0	1.5	0.4
200		0.7	ULSMV	-1.1	-0.6	-1.6	-0.2	0.4	-0.6	-0.7	-0.0	-1.1	-0.8	-0.0	-1.4	-0.1	0.3	-0.7	-1.1	-0.7	-1.6
200		0.7	WLSMV	0.5	1.2	-0.1	0.6	1.1	0.1	1.0	1.5	0.7	1.1	1.7	0.6	0.7	1.2	0.2	0.9	1.4	0.5
200		0.7	BAYES	0.7	2.4	-1.8	0.7	1.6	-1.4	0.8	2.4	-3.5	0.6	2.5	-16.6	0.7	2.1	-9.3	-0.0	2.1	-17.5
200	2 factors (φ = 0)	0.4	ML	-4.9	1.9	-9.1	-7.7	-4.7	-10.7	-3.9	-0.9	-7.2	-3.0	-0.8	-5.8	-8.3	-6.2	-9.9	-3.4	-0.1	-7.1
200		0.4	ULSMV	-6.9	5.1	-17.1	0.4	5.0	-2.3	-5.7	5.8	-14.4	-9.1	0.3	-13.4	-0.5	1.7	-2.1	-9.7	0.8	-14.0
200		0.4	WLSMV	-4.9	6.1	-13.4	1.0	5.9	-2.2	-3.6	6.6	-11.7	-10.7	1.5	-15.1	0.5	3.0	-1.1	-11.4	4.0	-16.9
200		0.4	BAYES	-13.4	6.6	-23.9	-2.9	8.4	-9.2	-13.3	8.0	-22.0	-4.4	1.1	-8.3	-1.2	2.5	-7.6	-5.2	0.5	-9.4
200		0.7	ML	-1.2	-0.3	-2.2	-3.5	-3.2	-3.9	-1.2	-0.8	-1.7	-0.1	0.6	-1.0	-3.5	-2.4	-4.3	-0.1	0.7	-0.7
200		0.7	ULSMV	-0.9	0.1	-1.8	-0.0	0.6	-0.4	-0.9	-0.2	-1.6	-1.0	-0.1	-2.1	-0.4	0.1	-1.0	-0.9	-0.1	-1.6
200		0.7	WLSMV	0.1	1.2	-0.7	0.4	1.1	0.0	0.3	1.1	-0.3	0.7	1.4	-0.2	0.3	0.8	-0.3	0.7	1.9	0.1
200		0.7	BAYES	-2.0	2.9	-16.9	-0.8	2.5	-9.7	-2.6	1.8	-16.1	-0.8	2.6	-18.5	-0.6	1.1	-12.1	-0.9	3.4	-17.2
200	2 factors (φ = 0.3)	0.4	ML	-5.1	0.5	-10.1	-7.2	-4.2	-9.9	-4.3	1.2	-7.5	-2.3	-0.4	-5.1	-8.7	-7.0	-10.1	-3.0	-1.4	-6.5
200		0.4	ULSMV	-7.5	5.7	-15.8	0.1	2.3	-3.4	-7.6	6.0	-16.1	-8.5	-0.3	-11.7	-1.2	0.1	-3.1	-9.3	1.5	-13.4
200		0.4	WLSMV	-5.1	11.6	-11.4	0.7	3.0	-2.9	-5.5	8.1	-16.5	-10.1	4.8	-15.8	-0.1	1.3	-2.1	-9.9	5.8	-15.5
200		0.4	BAYES	-12.5	7.5	-23.0	-1.8	8.2	-8.5	-13.2	8.7	-21.2	-3.7	1.0	-8.2	-1.2	2.1	-6.2	-4.5	2.0	-8.6
200		0.7	ML	-1.2	-0.5	-1.8	-3.8	-3.0	-4.5	-1.4	-0.8	-1.9	-0.2	0.4	-0.9	-3.3	-2.8	-3.7	0.1	0.6	-0.9
200		0.7	ULSMV	-1.1	-0.5	-1.7	-0.5	0.3	-1.0	-1.3	-0.6	-1.8	-1.1	-0.3	-1.8	-0.3	0.2	-0.9	-0.9	-0.3	-1.9
200		0.7	WLSMV	0.1	0.7	-0.4	0.0	0.7	-0.3	-0.0	0.5	-0.5	0.7	1.5	0.0	0.5	1.1	-0.1	0.9	1.5	-0.1
200		0.7	BAYES	-1.8	2.9	-15.6	-1.0	1.9	-8.5	-2.6	1.3	-14.9	-0.8	2.1	-17.6	-0.4	1.5	-11.8	-0.6	2.6	-16.6
200	2 factors (φ = 0.6)	0.4	ML	-3.7	0.7	-7.2	-7.7	-5.9	-9.2	-2.7	-0.8	-5.2	-2.1	0.6	-4.3	-8.3	-6.7	-11.5	-1.8	-0.2	-5.1
200		0.4	ULSMV	-6.4	2.5	-11.6	-0.4	0.7	-2.1	-4.7	4.7	-9.7	-8.9	-1.9	-12.4	-0.9	0.7	-3.4	-8.6	-1.1	-12.2
200		0.4	WLSMV	-3.6	4.4	-7.3	0.3	1.4	-1.2	-2.5	7.6	-9.4	-7.1	3.1	-11.7	0.3	2.0	-2.1	-6.9	3.9	-10.6
200		0.4	BAYES	-7.3	9.9	-16.4	1.5	10.2	-6.3	-6.8	11.3	-14.2	-1.3	4.0	-5.4	0.8	4.8	-2.6	-1.4	4.3	-6.9
200		0.7	ML	-1.0	-0.5	-1.5	-3.5	-3.0	-4.4	-0.9	-0.1	-1.9	0.0	0.5	-0.4	-3.1	-2.4	-4.2	0.1	0.9	-0.6
200		0.7	ULSMV	-1.4	-0.7	-2.1	-0.2	0.1	-0.6	-1.3	-0.1	-2.4	-1.3	-0.7	-2.0	-0.3	0.3	-1.0	-1.1	-0.3	-1.9
200		0.7	WLSMV	0.1	0.7	-0.5	0.4	0.8	0.0	0.3	1.4	-0.7	0.7	1.2	0.0	0.6	1.2	-0.1	0.8	1.6	0.1
200		0.7	BAYES	-1.0	3.1	-12.4	-0.4	2.1	-7.4	-1.6	2.0	-12.4	-0.4	2.4	-15.5	-0.1	1.5	-10.3	-0.3	2.9	-14.7
500	Unidimensional	0.4	ML	-0.9	1.5	-2.2	-8.9	-7.5	-9.8	-1.2	-0.3	-2.6	-0.1	1.3	-1.2	-8.7	-7.6	-9.7	-0.1	1.0	-1.2
500		0.4	ULSMV	-1.7	-0.2	-2.9	-0.4	0.4	-1.4	-1.9	-0.5	-3.7	-1.7	-0.2	-3.0	-0.3	0.2	-1.2	-1.8	-0.6	-2.8
500		0.4	WLSMV	0.3	2.4	-1.1	0.0	0.8	-1.0	-0.1	1.4	-1.5	0.4	1.7	-0.8	0.2	0.8	-0.7	0.3	1.4	-0.8
500		0.4	BAYES	-4.3	0.4	-24.6	-2.9	0.9	-21.6	-4.6	-0.6	-22.0	-3.6	1.4	-33.7	-2.4	1.1	-27.9	-3.4	2.1	-33.5
500		0.7	ML	0.0	0.5	-0.4	-3.3	-2.7	-3.8	-0.3	-0.0	-0.5	1.0	1.5	0.6	-2.8	-2.4	-3.3	1.0	1.4	0.6
500		0.7	ULSMV	-0.3	0.1	-0.7	0.0	0.4	-0.2	-0.5	-0.3	-0.7	-0.4	-0.0	-0.8	-0.1	0.1	-0.4	-0.4	-0.0	-0.8
500		0.7	WLSMV	0.4	0.8	-0.0	0.3	0.7	0.1	0.1	0.4	-0.1	0.4	0.7	-0.1	0.2	0.5	-0.1	0.4	0.8	-0.0
500		0.7	BAYES	0.3	1.7	-5.0	0.2	1.1	-4.0	-0.4	1.4	-5.8	-0.1	1.9	-12.4	0.1	1.1	-7.4	0.1	1.6	-11.6
500	2 factors (φ = 0)	0.4	ML	-1.5	1.1	-3.9	-7.8	-6.3	-9.3	-1.8	-0.2	-4.5	-1.0	0.3	-2.4	-8.6	-7.1	-9.7	-1.2	0.1	-2.6
500		0.4	ULSMV	-0.9	1.0	-3.4	0.4	1.4	-0.5	-0.8	1.6	-2.4	-1.9	-0.8	-3.6	-0.1	1.1	-1.1	-2.1	-0.8	-3.3
500		0.4	WLSMV	0.1	1.9	-2.6	0.7	1.6	-0.2	0.2	2.4	-1.7	0.0	1.4	-1.4	0.3	1.5	-0.6	-0.2	1.1	-1.7
500		0.4	BAYES	-2.4	3.4	-5.8	-0.5	2.0	-2.8	-3.1	0.7	-5.1	-1.6	0.8	-5.6	-0.8	1.5	-8.1	-2.0	1.7	-5.7
500		0.7	ML	-0.8	-0.2	-1.4	-3.6	-3.3	-4.0	-1.1	-0.3	-1.6	0.0	0.4	-0.4	-3.5	-2.7	-4.1	-0.2	0.2	-0.5
500		0.7	ULSMV	-0.2	0.5	-0.7	0.0	0.4	-0.2	-0.4	0.3	-0.8	-0.3	0.2	-0.7	-0.2	0.1	-0.5	-0.5	-0.1	-0.9
500		0.7	WLSMV	0.3	0.9	-0.2	0.2	0.7	-0.0	0.0	0.8	-0.5	0.4	0.9	0.0	0.1	0.4	-0.2	0.2	0.6	-0.2
500		0.7	BAYES	-0.7	1.5	-8.3	-0.3	1.8	-5.8	-1.2	1.8	-9.1	-0.2	1.6	-8.7	-0.3	1.0	-6.1	-0.6	1.3	-9.9
500	2 factors (φ = 0.3)	0.4	ML	-1.9	0.2	-5.4	-8.1	-6.5	-10.0	-1.0	1.0	-2.6	-1.0	1.6	-3.3	-8.6	-7.7	-11.1	-1.1	0.2	-2.9
500		0.4	ULSMV	-1.8	-0.2	-5.2	0.1	1.5	-1.4	-1.0	0.6	-2.3	-2.1	0.2	-4.8	-0.3	0.6	-2.4	-2.1	-0.8	-5.0
500		0.4	WLSMV	-0.5	1.2	-4.0	0.4	1.8	-1.1	0.1	2.0	-1.8	-0.2	2.3	-2.7	0.1	1.0	-1.9	-0.2	1.1	-2.5
500		0.4	BAYES	-1.9	4.8	-7.7	-0.0	2.1	-2.2	-1.4	2.8	-5.6	-1.2	1.8	-4.1	-0.6	2.1	-6.5	-1.4	1.7	-4.6
500		0.7	ML	-0.9	-0.4	-1.4	-3.6	-3.0	-4.3	-1.0	-0.4	-1.6	-0.1	0.4	-0.5	-3.3	-2.5	-3.9	0.0	0.4	-0.3
500		0.7	ULSMV	-0.3	0.2	-0.9	-0.0	0.2	-0.4	-0.4	-0.1	-1.1	-0.5	-0.1	-0.9	-0.1	0.2	-0.6	-0.3	0.1	-0.8
500		0.7	WLSMV	0.2	0.7	-0.3	0.2	0.4	-0.2	0.1	0.4	-0.5	0.2	0.6	-0.1	0.2	0.5	-0.3	0.4	0.8	-0.0
500		0.7	BAYES	-0.8	1.9	-7.7	-0.2	1.3	-5.5	-1.0	2.1	-9.0	-0.3	1.4	-8.6	-0.1	0.9	-6.0	-0.4	2.0	-9.6
500	2 factors (φ = 0.6)	0.4	ML	-1.5	0.4	-3.6	-8.8	-7.5	-10.5	-1.2	-0.2	-2.4	-0.3	1.9	-1.5	-8.5	-6.8	-9.8	-0.5	1.4	-2.1
500		0.4	ULSMV	-2.3	-0.0	-5.1	-0.7	0.1	-1.6	-1.8	-0.4	-2.8	-2.0	0.6	-3.3	-0.3	0.5	-1.2	-2.0	-1.0	-4.3
500		0.4	WLSMV	-0.8	0.6	-3.3	-0.3	0.4	-1.2	-0.2	0.6	-1.5	0.1	2.4	-1.1	0.2	1.0	-0.6	0.1	1.6	-1.8
500		0.4	BAYES	0.9	8.8	-2.5	1.3	6.4	-1.0	1.1	6.6	-3.2	0.7	2.2	-1.6	0.6	3.9	-1.7	0.6	3.2	-4.1
500		0.7	ML	-0.7	-0.1	-1.1	-3.7	-3.0	-4.3	-0.9	-0.2	-1.3	0.1	0.6	-0.3	-3.2	-2.6	-4.0	0.1	0.5	-0.2
500		0.7	ULSMV	-0.4	0.0	-0.8	-0.2	0.2	-0.5	-0.5	0.1	-1.0	-0.5	0.1	-1.0	-0.1	0.3	-0.3	-0.5	0.0	-0.9
500		0.7	WLSMV	0.2	0.7	-0.2	0.1	0.5	-0.3	0.1	0.6	-0.5	0.3	0.8	-0.2	0.3	0.7	-0.0	0.3	0.8	-0.1
500		0.7	BAYES	-0.5	1.5	-6.6	-0.2	1.2	-5.1	-0.8	2.0	-7.7	-0.1	1.8	-7.2	-0.0	1.1	-5.1	-0.3	1.7	-7.9
1000	Unidimensional	0.4	ML	-0.8	1.1	-1.5	-8.8	-7.9	-9.8	-0.8	0.1	-1.9	0.1	0.9	-1.0	-8.6	-7.4	-9.6	-0.1	0.9	-0.7
1000		0.4	ULSMV	-0.9	0.5	-1.8	-0.1	0.4	-0.8	-0.9	-0.3	-1.9	-0.7	0.0	-1.9	-0.1	0.7	-0.9	-0.9	-0.0	-1.6
1000		0.4	WLSMV	0.0	1.7	-0.8	0.1	0.7	-0.5	0.0	0.6	-1.1	0.3	1.0	-0.8	0.2	1.0	-0.7	0.2	1.0	-0.5
1000		0.4	BAYES	-2.3	2.2	-15.2	-1.3	0.9	-7.7	-2.6	0.0	-16.4	-1.8	2.1	-27.8	-1.1	1.6	-19.8	-1.7	1.3	-26.5
1000		0.7	ML	-0.0	0.2	-0.2	-3.4	-3.1	-4.1	-0.1	0.2	-0.4	1.1	1.4	0.8	-2.8	-2.4	-3.4	1.1	1.3	0.8
1000		0.7	ULSMV	-0.1	0.1	-0.3	-0.0	0.2	-0.2	-0.2	0.1	-0.5	-0.1	0.1	-0.5	-0.1	0.1	-0.3	-0.1	0.2	-0.4
1000		0.7	WLSMV	0.2	0.5	0.1	0.1	0.3	-0.1	0.1	0.4	-0.2	0.2	0.5	-0.1	0.1	0.2	-0.1	0.2	0.5	-0.0
1000		0.7	BAYES	0.0	1.1	-3.5	0.0	0.6	-0.6	-0.0	0.8	-3.4	0.1	0.8	-6.4	0.1	0.7	-3.5	0.1	1.0	-6.2
1000	2 factors (φ	0.4	ML	-1.1	1.0	-2.4	-8.5	-7.2	-9.6	-1.4	0.4	-2.7	-0.7	0.6	-1.9	-8.6	-7.3	-9.4	-0.8	0.5	-2.0
1000		0.4	ULSMV	-0.6	0.8	-1.9	0.1	1.1	-1.1	-0.8	0.2	-1.9	-0.8	0.6	-2.5	0.1	0.9	-0.5	-0.8	0.2	-2.1
1000		0.4	WLSMV	0.0	1.8	-1.5	0.2	1.2	-1.0	-0.2	1.0	-1.2	0.1	1.6	-1.3	0.3	1.1	-0.3	0.1	1.1	-1.1
1000		0.4	BAYES	-1.2	2.9	-8.1	-0.4	2.2	-1.9	-1.5	1.1	-3.2	-0.8	1.6	-3.9	-0.4	1.4	-4.7	-0.9	3.1	-2.9

Sample Size	Model	MFL	Method	Number of Items = 10									Number of Items = 20								
				Left-Skewed			Normal			Right-Skewed			Left-Skewed			Normal			Right-Skewed		
				Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
1000		0.7	ML	-0.9	-0.8	-1.0	-3.7	-3.1	-4.1	-1.0	-0.7	-1.2	-0.1	0.3	-0.3	-3.4	-2.8	-3.9	-0.1	0.2	-0.6
1000		0.7	ULSMV	-0.1	0.0	-0.2	-0.0	0.2	-0.2	-0.2	0.1	-0.5	-0.1	0.2	-0.4	-0.1	0.1	-0.3	-0.2	0.1	-0.6
1000		0.7	WLSMV	0.1	0.3	-0.1	0.1	0.3	-0.1	-0.0	0.3	-0.3	0.2	0.5	-0.0	0.1	0.2	-0.1	0.1	0.4	-0.3
1000		0.7	BAYES	-0.3	1.8	-5.3	-0.1	0.9	-2.5	-0.5	1.1	-5.1	-0.1	1.0	-3.1	-0.1	0.4	-2.6	-0.3	1.2	-3.0
1000	2 factors (φ = 0.3)	0.4	ML	-1.3	-0.1	-2.4	-8.6	-7.3	-9.6	-1.1	-0.4	-2.8	-0.7	1.2	-1.6	-8.6	-7.3	-9.9	-0.9	0.0	-1.8
1000		0.4	ULSMV	-0.9	0.2	-1.8	-0.0	1.2	-0.9	-0.7	0.4	-2.0	-1.0	0.8	-2.4	-0.1	0.5	-0.9	-1.2	0.2	-2.4
1000		0.4	WLSMV	-0.2	0.6	-0.9	0.1	1.4	-0.8	-0.0	1.0	-1.3	-0.0	1.9	-1.1	0.2	0.8	-0.7	-0.2	0.8	-1.3
1000		0.4	BAYES	-0.6	5.0	-6.6	0.0	2.7	-2.9	-0.4	2.8	-2.3	-0.6	1.8	-3.9	-0.3	1.4	-3.4	-0.9	2.0	-3.2
1000		0.7	ML	-0.9	-0.6	-1.2	-3.7	-3.2	-4.3	-0.9	-0.4	-1.1	-0.0	0.4	-0.3	-3.4	-2.6	-4.0	-0.1	0.2	-0.3
1000		0.7	ULSMV	-0.2	0.0	-0.5	-0.1	0.1	-0.4	-0.2	0.3	-0.6	-0.2	0.3	-0.6	-0.0	0.2	-0.2	-0.2	0.0	-0.6
1000		0.7	WLSMV	0.1	0.4	-0.2	-0.0	0.2	-0.3	0.1	0.6	-0.3	0.1	0.7	-0.1	0.1	0.3	-0.0	0.1	0.4	-0.1
1000		0.7	BAYES	-0.3	1.7	-5.1	-0.2	0.9	-2.8	-0.4	1.2	-5.2	-0.1	1.2	-3.3	-0.0	0.4	-2.5	-0.3	1.1	-3.4
1000	2 factors (φ = 0.6)	0.4	ML	-1.3	0.3	-1.9	-8.7	-8.1	-9.7	-1.0	0.0	-3.3	-0.4	0.5	-1.4	-8.6	-7.8	-9.5	-0.5	0.6	-1.2
1000		0.4	ULSMV	-1.1	0.2	-1.8	-0.2	0.4	-0.9	-0.9	0.2	-2.9	-1.1	-0.0	-1.8	-0.1	1.2	-1.1	-1.1	-0.1	-1.6
1000		0.4	WLSMV	-0.4	1.0	-1.0	0.0	0.6	-0.7	-0.1	0.9	-2.2	-0.0	0.8	-0.9	0.1	1.5	-0.9	-0.0	1.0	-0.6
1000		0.4	BAYES	1.5	10.2	-5.0	1.2	6.5	-1.9	1.9	8.0	-2.4	0.6	4.8	-2.5	0.3	1.9	-1.2	0.4	4.3	-2.0
1000		0.7	ML	-0.6	-0.4	-1.0	-3.6	-3.2	-4.0	-0.8	-0.4	-1.0	0.2	0.3	-0.1	-3.4	-2.8	-3.9	-0.0	0.4	-0.4
1000		0.7	ULSMV	-0.1	0.1	-0.6	-0.0	0.1	-0.1	-0.2	0.3	-0.7	-0.2	0.1	-0.4	-0.1	0.0	-0.4	-0.3	-0.0	-0.7
1000		0.7	WLSMV	0.1	0.4	-0.2	0.1	0.3	0.0	0.0	0.5	-0.3	0.2	0.5	-0.0	0.0	0.2	-0.2	0.0	0.4	-0.3
1000		0.7	BAYES	-0.1	1.9	-4.2	-0.0	1.0	-2.1	-0.3	1.2	-4.5	0.0	1.1	-2.7	-0.1	0.4	-2.6	-0.3	1.0	-3.2

Appendix 4. Relative percentage bias of interfactor correlations

Model	Sample Size	Number of Items	Estimation Method	Mean Factor Loading = 0.4			Mean Factor Loading = 0.7		
				Left-Skewed	Normal	Right-Skewed	Left-Skewed	Normal	Right-Skewed
2 factors (φ = 0.3)	200	10 Items	ML	-11.77	-2.13	-9.51	1.21	0.37	3.01
		10 Items	ULSMV	-27.72	7.51	-32.22	-4.45	1.88	-2.22
		10 Items	WLSMV	-9.93	8.88	-14.92	-1.24	2.55	0.59
		10 Items	BAYES	-52.33	-29.60	-53.21	-0.84	2.69	1.93
		20 Items	ML	-1.83	0.77	-2.25	4.63	1.06	0.47
		20 Items	ULSMV	-19.62	10.02	-19.00	-0.52	1.87	-5.39
		20 Items	WLSMV	-15.13	10.70	-14.29	3.41	2.49	-1.75
		20 Items	BAYES	-34.01	-19.16	-34.66	0.31	-1.17	-3.25
	500	10 Items	ML	-0.35	1.87	-1.81	1.78	0.31	1.91
		10 Items	ULSMV	-4.83	5.29	-5.98	-1.78	0.77	-1.64
		10 Items	WLSMV	2.99	5.56	2.75	0.86	1.01	0.97
		10 Items	BAYES	-27.78	-15.68	-29.73	0.39	0.25	0.76
		20 Items	ML	3.31	0.55	1.75	4.04	1.21	3.93
		20 Items	ULSMV	-1.72	4.45	-3.41	-0.46	0.89	-0.64
		20 Items	WLSMV	7.10	4.65	5.40	2.32	1.13	2.09
		20 Items	BAYES	-17.17	-9.67	-18.84	-0.36	-0.39	-0.65
	1000	10 Items	ML	2.75	0.99	2.87	2.60	0.44	2.78
		10 Items	ULSMV	-0.50	2.39	-0.68	-0.50	0.38	-0.34
		10 Items	WLSMV	3.68	2.52	3.59	0.78	0.49	0.96
		10 Items	BAYES	-16.36	-4.71	-14.68	0.31	0.80	0.80
		20 Items	ML	2.39	0.35	0.22	3.95	1.50	3.97
		20 Items	ULSMV	-0.97	2.36	-3.38	-0.13	0.81	0.26
		20 Items	WLSMV	3.37	2.46	0.94	1.21	0.92	1.58
		20 Items	BAYES	-9.30	-3.44	-10.70	0.03	0.44	-0.11
2 factors (φ = 0.6)	200	10 Items	ML	-7.98	0.24	-6.08	3.56	0.30	2.05
		10 Items	ULSMV	-17.69	1.78	-17.11	1.70	0.75	0.04
		10 Items	WLSMV	-11.77	2.12	-9.89	3.79	1.33	2.34
		10 Items	BAYES	-54.18	-31.97	-51.69	-1.14	0.28	-1.18
		20 Items	ML	4.01	-0.30	0.20	2.94	1.00	2.46
		20 Items	ULSMV	-1.78	4.49	-3.28	0.15	1.06	-0.33
		20 Items	WLSMV	2.84	4.83	2.50	2.54	1.54	2.06
		20 Items	BAYES	-32.52	-22.84	-35.04	-2.30	-1.53	-2.79
	500	10 Items	ML	1.79	0.63	-0.45	2.18	0.41	2.21
		10 Items	ULSMV	-0.76	2.35	-2.84	0.15	0.43	0.35
		10 Items	WLSMV	1.72	2.53	-0.08	1.08	0.63	1.29
		10 Items	BAYES	-29.27	-20.40	-30.88	0.35	-0.05	0.44
		20 Items	ML	2.12	0.39	1.66	2.77	1.00	3.04
		20 Items	ULSMV	1.37	2.31	0.79	-0.21	0.62	0.29
		20 Items	WLSMV	3.63	2.44	3.14	0.76	0.81	1.24
		20 Items	BAYES	-19.28	-11.67	-19.67	-0.99	-0.30	-0.78
	1000	10 Items	ML	1.60	0.17	2.35	2.34	0.48	1.80
		10 Items	ULSMV	0.85	0.97	1.60	0.22	0.23	-0.40
		10 Items	WLSMV	2.07	1.07	2.77	0.73	0.33	0.07
		10 Items	BAYES	-19.65	-10.51	-18.50	0.38	0.38	0.01
		20 Items	ML	1.24	-0.22	1.95	2.87	0.66	2.38
		20 Items	ULSMV	0.26	0.74	0.91	0.26	0.14	-0.19
		20 Items	WLSMV	1.46	0.81	2.10	0.72	0.23	0.29
		20 Items	BAYES	-11.40	-6.10	-10.66	-0.21	-0.23	-0.70

Appendix 5. Mean, maximum and minimum values of coverage rate

Sample Size	Model	MFL	Method	Number of Items = 10									Number of Items = 20								
				Left-Skewed			Normal			Right-Skewed			Left-Skewed			Normal			Right-Skewed		
				Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
200	Unidimensional	0.4	BAYES	92.5	94.6	91.0	93.8	95.7	87.6	92.9	94.1	91.9	91.1	95.3	38.0	91.0	95.9	31.3	91.2	95.8	36.6
200		0.4	ML	91.7	93.6	87.1	93.0	93.8	91.5	92.0	93.5	88.3	93.2	94.4	91.3	92.8	94.6	90.5	93.6	95.6	91.6
200		0.4	ULSMV	88.5	92.1	86.6	93.5	94.2	92.9	89.0	93.9	85.7	91.4	94.6	88.8	93.9	95.7	92.8	91.4	94.3	88.9
200		0.4	WLSMV	87.8	90.2	86.3	92.8	93.4	92.2	88.0	91.3	85.8	89.0	93.2	86.4	93.1	94.7	91.6	88.9	91.5	86.6
200		0.7	BAYES	91.1	94.3	88.7	92.6	94.3	91.3	91.8	92.7	90.6	89.7	93.0	57.5	91.4	94.1	78.3	89.7	94.3	55.3
200		0.7	ML	93.5	95.3	92.8	94.3	95.6	93.4	93.1	95.0	92.2	92.0	93.8	90.1	94.5	96.2	93.1	92.2	93.3	91.0
200		0.7	ULSMV	93.2	94.2	92.4	94.0	95.3	92.5	92.7	93.7	91.8	92.9	94.1	91.7	94.1	95.1	92.9	93.0	94.5	91.5
200		0.7	WLSMV	91.1	91.6	90.4	93.0	94.3	91.9	90.6	91.7	89.4	90.3	91.7	89.2	92.8	93.6	91.8	90.7	92.9	88.6
200	2 factors(φ = 0)	0.4	BAYES	90.9	99.7	87.8	93.8	99.9	90.6	91.4	99.6	88.2	93.2	100.0	90.7	93.6	98.3	91.2	93.1	100.0	89.5
200		0.4	ML	84.0	87.5	81.6	89.2	91.1	87.7	84.7	87.5	80.7	91.4	93.3	88.1	92.6	94.4	90.3	91.2	93.0	86.7
200		0.4	ULSMV	88.3	96.7	85.1	92.0	94.4	90.6	87.7	96.2	84.4	87.0	93.1	84.5	92.8	94.6	91.0	86.5	93.2	83.9
200		0.4	WLSMV	86.8	93.5	84.1	91.8	94.3	90.1	87.3	93.9	83.4	79.1	85.0	74.4	92.1	94.2	90.5	79.0	84.4	76.1
200		0.7	BAYES	87.1	92.1	74.3	90.1	92.8	81.6	88.3	92.5	74.6	89.4	95.0	53.8	90.7	94.7	71.9	89.7	94.0	62.6
200		0.7	ML	93.1	94.0	92.5	94.5	95.5	93.4	93.2	94.2	91.9	93.0	95.4	91.8	94.0	95.6	92.6	93.3	94.1	91.7
200		0.7	ULSMV	92.4	93.4	91.4	94.0	95.0	92.7	92.8	94.2	91.4	92.7	94.5	91.2	93.9	95.3	92.4	93.2	94.5	92.3
200		0.7	WLSMV	90.8	92.3	89.9	93.4	95.1	92.2	91.1	92.0	89.5	89.4	91.8	87.4	93.0	94.7	91.3	90.2	91.0	89.0
200	2 factors(φ = 0.3)	0.4	BAYES	91.4	100.0	87.0	93.5	99.4	89.7	91.0	99.6	87.5	93.0	100.0	90.5	93.8	98.3	91.9	93.0	100.0	88.9
200		0.4	ML	85.5	88.2	82.3	90.0	91.5	88.6	85.7	88.3	83.0	91.4	92.8	88.9	92.5	94.3	90.9	91.5	93.1	88.5
200		0.4	ULSMV	87.8	95.6	83.5	92.3	93.2	91.6	87.5	95.1	83.3	87.2	94.5	84.8	92.9	94.4	91.1	87.4	93.9	83.9
200		0.4	WLSMV	87.5	93.6	84.3	91.9	92.8	91.0	87.4	94.2	83.9	82.2	89.5	79.3	92.0	93.7	90.4	82.8	89.4	79.9
200		0.7	BAYES	87.9	92.9	74.8	90.5	93.7	86.7	88.6	92.3	78.2	89.5	93.8	58.1	90.8	94.0	72.7	89.4	93.2	62.8
200		0.7	ML	93.8	94.6	92.5	93.9	94.8	92.7	93.5	95.1	92.2	93.0	94.6	91.9	94.1	95.8	92.8	92.8	94.5	90.5
200		0.7	ULSMV	92.9	94.1	91.4	93.7	95.2	92.0	92.8	94.8	91.8	92.7	93.9	91.5	93.9	94.7	92.8	92.5	93.4	90.6
200		0.7	WLSMV	92.0	93.1	90.8	93.0	94.2	91.4	91.7	94.3	90.5	90.4	92.0	89.3	92.8	94.0	91.6	90.2	91.2	88.3
200	2 factors(φ = 0.6)	0.4	BAYES	91.0	99.3	87.3	93.7	98.5	91.3	90.8	99.3	87.8	93.1	100.0	90.5	93.6	98.9	90.0	93.6	100.0	89.6
200		0.4	ML	88.1	91.0	85.6	91.8	92.8	91.0	88.3	89.6	86.4	92.2	94.3	89.7	92.7	94.4	91.0	92.7	94.8	89.3
200		0.4	ULSMV	88.8	93.9	85.8	92.9	94.0	91.8	88.3	93.4	85.3	88.7	95.2	86.5	93.4	94.6	92.2	88.7	94.2	86.4
200		0.4	WLSMV	88.1	91.9	83.7	92.5	93.5	91.2	88.2	92.7	84.9	85.3	92.4	83.2	92.6	94.1	91.3	85.7	91.6	83.0
200		0.7	BAYES	89.0	91.8	80.6	90.7	92.9	88.3	89.7	92.4	79.1	90.1	93.8	65.8	91.2	93.8	79.2	90.0	93.5	68.9
200		0.7	ML	93.3	94.4	92.0	93.8	95.4	92.5	93.0	93.8	92.3	92.9	94.5	91.1	94.4	95.1	93.7	93.1	94.1	90.7
200		0.7	ULSMV	93.1	94.2	91.8	93.5	94.8	92.2	92.7	93.6	92.0	92.8	94.0	91.4	94.1	95.5	93.0	93.0	94.6	91.3
200		0.7	WLSMV	91.6	92.8	90.1	92.7	93.9	91.6	91.3	92.9	90.2	90.9	93.2	89.1	92.9	94.4	91.0	90.8	93.0	88.5
500	Unidimensional	0.4	BAYES	88.7	94.5	46.8	89.6	96.1	46.0	90.8	95.7	64.1	90.1	96.4	14.3	90.7	95.9	26.2	90.0	95.6	15.0
500		0.4	ML	94.2	95.8	93.2	92.1	93.5	90.0	94.3	95.3	92.4	94.1	95.4	92.7	90.7	91.9	89.1	94.3	95.7	92.4
500		0.4	ULSMV	93.0	94.7	91.9	94.6	95.6	93.8	93.3	94.7	91.7	93.4	94.6	91.8	94.5	95.7	93.1	93.5	95.1	91.8
500		0.4	WLSMV	92.9	94.7	91.7	94.3	95.4	93.7	93.0	94.0	92.1	93.2	94.1	91.5	94.1	95.3	92.5	93.1	94.6	91.7
500		0.7	BAYES	91.5	93.6	82.8	91.8	95.0	79.8	92.4	94.5	87.0	90.6	94.9	52.3	92.5	96.3	68.5	91.1	94.3	57.2
500		0.7	ML	94.3	95.7	93.4	93.4	95.2	92.3	94.6	95.5	93.8	93.2	94.7	91.7	93.6	94.9	92.7	93.5	95.0	91.7
500		0.7	ULSMV	94.2	95.5	93.2	94.5	96.1	93.4	94.4	95.2	93.7	93.9	95.4	93.0	94.9	96.1	93.7	94.4	95.5	92.9
500		0.7	WLSMV	93.6	95.5	92.3	94.0	95.5	92.6	93.6	94.5	93.0	93.0	94.4	91.8	94.4	96.0	92.7	93.5	94.9	92.2
500	2 factors(φ = 0)	0.4	BAYES	94.0	99.9	90.8	94.4	98.3	92.8	94.2	99.8	91.5	94.0	99.6	92.3	94.2	98.9	85.3	94.1	99.6	92.1
500		0.4	ML	92.5	93.8	90.1	92.2	93.7	90.5	92.9	94.8	91.6	94.3	96.4	92.8	91.6	93.1	90.2	94.2	95.6	92.5
500		0.4	ULSMV	91.5	93.8	90.6	93.9	94.7	93.2	92.5	94.0	90.5	92.9	95.3	91.1	94.3	95.3	92.9	92.8	94.3	90.9
500		0.4	WLSMV	92.0	94.0	91.0	93.8	94.7	92.9	92.9	94.6	91.5	92.8	94.7	90.8	94.0	95.1	92.6	92.7	93.9	91.0
500		0.7	BAYES	91.3	94.3	84.0	91.8	93.8	87.0	90.6	93.7	80.6	91.9	94.6	78.2	92.3	94.3	77.8	92.3	95.3	75.9
500		0.7	ML	94.4	95.6	93.6	92.9	95.7	91.2	94.5	95.8	92.8	94.2	96.3	92.7	93.0	95.1	91.5	94.3	95.3	93.5
500		0.7	ULSMV	93.8	95.0	92.3	94.1	95.7	93.2	93.9	95.5	92.1	94.1	95.3	92.2	94.5	95.1	93.6	94.0	95.3	92.5
500		0.7	WLSMV	93.7	94.9	92.0	93.9	95.5	92.9	93.8	95.3	92.1	93.3	95.1	91.7	94.1	95.0	93.2	93.3	94.5	91.9
500	2 factors(φ = 0.3)	0.4	BAYES	94.3	99.7	91.4	94.5	98.4	91.9	94.5	99.7	91.2	93.9	99.8	91.0	94.0	98.6	89.2	94.6	99.8	92.3
500		0.4	ML	93.3	94.8	91.7	92.4	93.8	91.3	94.3	95.0	93.8	94.2	95.4	92.9	92.0	93.1	89.9	94.4	96.2	92.8
500		0.4	ULSMV	92.3	93.2	91.4	93.8	94.3	92.5	93.1	93.8	92.2	92.7	94.3	91.2	94.4	95.6	93.4	93.0	94.8	91.3
500		0.4	WLSMV	92.5	93.6	91.3	93.6	94.4	92.4	93.1	94.2	91.7	92.6	94.6	91.4	94.0	95.4	93.0	92.9	94.3	91.7
500		0.7	BAYES	91.0	93.7	84.5	92.4	94.5	88.9	90.1	93.7	79.1	92.4	95.6	79.0	92.7	95.3	77.9	92.0	95.1	74.9
500		0.7	ML	94.2	95.3	92.9	93.3	94.9	91.8	94.3	95.3	93.3	94.6	95.6	93.1	93.4	94.6	92.1	94.1	95.7	92.7
500		0.7	ULSMV	94.0	94.9	92.8	94.6	95.6	92.2	94.1	94.7	93.0	94.2	95.8	92.5	94.7	96.0	93.2	93.8	95.1	92.8
500		0.7	WLSMV	93.4	94.8	92.7	94.3	95.9	92.4	93.5	94.6	92.2	93.7	95.1	92.5	94.3	95.9	93.0	93.0	94.4	91.4
500	2 factors(φ = 0.6)	0.4	BAYES	94.3	97.9	91.6	94.6	98.7	92.9	94.0	98.6	90.6	94.1	99.3	92.0	94.2	96.1	92.5	94.3	99.4	91.9
500		0.4	ML	94.3	95.0	93.4	92.6	93.1	91.8	93.8	95.4	92.1	94.3	95.7	93.3	91.8	93.8	90.1	94.3	95.4	92.5
500		0.4	ULSMV	92.8	93.7	91.7	94.3	95.4	93.2	92.3	94.4	91.4	93.6	94.8	92.4	94.5	95.6	93.5	93.2	94.4	91.1
500		0.4	WLSMV	92.7	93.6	91.8	94.2	95.4	93.0	92.4	94.4	91.5	93.2	94.3	91.9	94.1	95.4	93.3	93.0	94.5	90.9
500		0.7	BAYES	91.9	94.2	86.0	92.7	95.7	88.6	91.6	94.9	80.7	92.9	96.0	84.0	92.9	94.7	81.2	93.0	95.6	82.4
500		0.7	ML	94.7	96.3	93.0	93.2	94.7	92.1	94.4	95.5	92.8	94.7	96.7	91.8	93.7	95.2	92.2	94.5	95.6	93.4
500		0.7	ULSMV	94.3	95.7	93.4	94.8	96.0	93.6	94.0	95.5	92.8	94.7	96.4	92.9	95.0	95.9	93.9	94.3	95.3	93.1
500		0.7	WLSMV	93.7	95.1	92.5	94.5	95.8	93.5	93.6	95.2	92.2	93.8	95.6	91.9	94.4	95.5	93.4	93.6	95.4	92.3
1000	Unidimensional	0.4	BAYES	90.6	94.9	63.8	91.8	95.6	71.4	90.9	94.8	65.4	91.0	96.7	23.0	91.2	95.5	37.5	90.6	95.6	25.5
1000		0.4	ML	94.4	95.6	92.8	89.1	90.8	86.9	94.8	95.6	93.5	95.1	96.2	94.2	87.2	89.2	84.6	94.7	95.8	93.7
1000		0.4	ULSMV	93.7	94.6	92.4	94.4	96.4	93.7	94.1	95.0	92.5	94.7	95.9	93.2	94.6	95.6	93.6	94.2	95.6	93.0
1000		0.4	WLSMV	93.6	94.8	92.4	94.4	96.4	93.6	94.0	94.6	92.7	94.5	95.6	93.4	94.4	95.3	93.2	93.9	95.6	92.3
1000		0.7	BAYES	91.6	94.9	82.3	93.2	95.7	87.5	92.1	95.7	82.6	92.2	94.8	72.5	92.8	95.3	74.0	92.6	94.9	74.9
1000		0.7	ML	94.5	95.8	93.3	89.8	91.3	88.1	94.7	95.3	94.1	93.1	94.7	91.8	90.6	92.2	88.8	93.5	95.1	92.0
1000		0.7	ULSMV	94.3	95.3	93.7	94.6	95.5	93.1	94.3	95.0	93.9	94.2	95.2	92.1	94.8	96.2	93.6	94.8	96.0	93.6
1000		0.7	WLSMV	93.9	95.4	93.2	94.4	95.3	93.1	94.0	94.9	93.5	93.7	94.9	91.3	94.6	96.0	93.2	94.2	95.2	93.4
1000	2 factors(φ	0.4	BAYES	94.7	98.0	93.0	94.0	96.2	92.2	94.5	98.3	91.9	94.0	96.8	91.8	94.4	96.3	88.3	94.1	97.3	89.5
1000		0.4	ML	94.7	96.3	93.4	90.6	91.6	89.8	94.7	95.8	93.6	94.6	95.6	93.3	89.8	90.4	88.3	94.8	95.9	93.8
1000		0.4	ULSMV	93.3	94.2	92.0	94.1	95.1	92.1	93.6	95.2	92.1	93.9	95.1	92.5	94.8	96.5	93.2	94.3	96.2	93.1
1000		0.4	WLSMV	93.5	94.6	92.1	94.0	95.2	92.2	93.7	95.3	92.4	93.9	95.1	92.6	94.7	96.4	93.3	94.1	95.7	93.0

Sample Size	Model	MFL	Method	Number of Items = 10									Number of Items = 20								
				Left-Skewed			Normal			Right-Skewed			Left-Skewed			Normal			Right-Skewed		
				Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
1000		0.7	BAYES	92.4	95.0	84.2	93.4	95.9	90.1	92.3	94.3	87.6	92.4	96.1	82.3	92.8	94.8	84.9	93.3	96.6	85.0
1000		0.7	ML	95.2	95.9	94.4	90.7	92.6	89.3	94.7	95.7	93.1	94.9	95.8	93.7	89.5	91.0	87.6	95.0	96.1	94.0
1000		0.7	ULSMV	94.7	95.7	93.7	95.1	96.3	94.4	94.2	95.2	93.0	94.5	95.5	93.2	94.4	95.2	93.1	94.8	96.0	93.7
1000		0.7	WLSMV	94.6	95.5	94.0	95.1	96.2	94.5	94.4	95.3	93.7	94.2	95.3	93.0	94.2	95.1	93.2	94.4	96.2	92.8
1000	2 factors(φ = 0.3)	0.4	BAYES	94.6	97.5	93.0	94.5	96.5	92.5	94.9	98.6	92.5	93.7	96.8	91.7	94.4	96.1	90.3	94.1	98.6	91.1
1000		0.4	ML	94.6	95.8	93.6	90.8	92.8	88.7	95.0	95.8	94.5	94.6	95.8	93.5	89.3	91.0	87.7	94.7	95.9	93.4
1000		0.4	ULSMV	93.5	94.9	92.2	94.5	95.4	93.3	94.1	95.2	93.2	94.0	94.7	93.3	94.7	96.0	93.7	94.2	95.8	92.3
1000		0.4	WLSMV	93.5	94.9	91.9	94.4	95.3	93.2	94.0	95.1	93.3	93.8	95.0	92.9	94.6	96.0	93.5	94.0	95.4	92.2
1000		0.7	BAYES	92.3	96.6	86.8	93.3	95.2	87.2	92.6	95.0	86.1	92.7	95.4	82.1	92.9	95.8	85.5	92.9	96.5	84.8
1000		0.7	ML	94.9	96.1	94.1	90.3	92.9	88.7	94.9	96.0	93.6	95.0	96.2	93.4	89.3	91.2	87.7	94.7	96.5	92.8
1000		0.7	ULSMV	94.6	96.2	94.0	94.7	95.9	93.1	94.8	95.6	94.0	94.7	95.8	93.4	94.6	95.8	93.3	94.5	96.4	92.6
1000		0.7	WLSMV	94.4	95.6	93.8	94.8	96.0	93.3	94.5	95.2	93.6	94.3	95.8	93.0	94.2	95.5	93.4	94.2	96.5	92.7
1000	2 factors(φ = 0.6)	0.4	BAYES	94.2	98.4	90.8	94.5	95.7	93.6	94.0	99.3	91.3	94.0	96.8	91.9	94.8	96.4	93.2	94.5	97.6	91.7
1000		0.4	ML	94.5	95.3	93.4	90.5	91.3	89.5	94.5	95.5	92.8	94.8	96.0	93.2	88.9	90.5	86.8	95.0	96.0	93.4
1000		0.4	ULSMV	93.7	94.3	92.7	94.6	95.9	93.5	93.5	95.4	91.8	94.2	95.2	92.8	94.7	95.5	93.6	94.5	95.5	93.1
1000		0.4	WLSMV	93.6	94.5	92.8	94.6	95.7	93.5	93.5	95.5	92.0	94.1	95.5	92.7	94.5	95.2	93.6	94.4	95.3	93.0
1000		0.7	BAYES	93.0	95.4	87.5	93.6	96.6	90.3	92.4	95.1	87.1	92.5	95.5	85.1	93.2	95.8	85.3	93.3	95.9	84.9
1000		0.7	ML	94.9	95.9	93.5	90.9	92.2	89.4	95.0	96.5	92.6	94.6	95.4	93.0	89.3	91.0	87.7	94.8	95.7	93.4
1000		0.7	ULSMV	94.8	96.5	93.0	95.3	96.3	94.3	94.4	95.6	92.6	94.8	96.2	93.6	94.7	96.2	93.6	94.5	95.7	92.9
1000		0.7	WLSMV	94.3	95.3	92.8	94.9	96.0	93.9	94.3	95.3	92.1	94.3	95.2	92.8	94.5	96.2	93.5	94.3	95.6	92.8

Appendix 6. Coverage rate of interfactor correlations

Sample Size	Model	Mean Factor Loading	Estimation Method	Number of Items = 10			Number of Items = 20		
				Left-Skewed	Normal	Right-Skewed	Left-Skewed	Normal	Right-Skewed
200	2 factors (φ = 0)	0.4	ML	76.98	88.48	80.63	90.60	92.70	90.20
		0.4	ULSMV	71.31	84.82	71.86	63.44	86.86	61.01
		0.4	WLSMV	67.74	83.54	70.81	38.22	85.74	40.11
		0.4	BAYES	96.79	95.10	96.90	97.09	95.10	96.49
		0.7	ML	93.00	94.60	95.10	94.50	95.30	94.60
		0.7	ULSMV	86.80	93.30	89.00	86.90	93.60	87.30
		0.7	WLSMV	74.60	92.60	74.77	67.00	92.90	70.10
		0.7	BAYES	93.30	93.80	94.90	94.20	94.40	94.50
	2 factors (φ = 0.3)	0.4	ML	76.86	90.09	79.57	89.10	93.10	91.30
		0.4	ULSMV	72.52	87.66	71.77	73.36	88.03	75.97
		0.4	WLSMV	74.69	86.84	75.63	58.20	87.22	59.37
		0.4	BAYES	92.70	93.60	92.49	92.30	94.20	93.70
		0.7	ML	92.40	94.00	93.10	93.50	94.50	93.10
		0.7	ULSMV	89.10	93.20	89.20	92.00	94.50	89.40
		0.7	WLSMV	86.29	92.20	86.29	86.90	93.60	85.50
		0.7	BAYES	92.80	93.50	93.40	94.40	95.10	93.70
	2 factors (φ = 0.6)	0.4	ML	78.68	88.64	80.73	90.78	92.59	90.29
		0.4	ULSMV	84.82	94.21	85.51	84.63	91.04	83.37
		0.4	WLSMV	86.86	93.76	87.37	79.42	90.13	80.30
		0.4	BAYES	71.57	83.30	75.20	77.15	81.10	71.64
		0.7	ML	92.60	94.10	92.50	91.80	94.40	92.50
		0.7	ULSMV	91.60	93.40	91.50	92.20	93.40	91.40
		0.7	WLSMV	90.70	93.00	90.80	90.50	92.60	89.59
		0.7	BAYES	93.70	94.60	94.40	94.30	95.60	95.30
500	2 factors (φ = 0)	0.4	ML	90.69	94.00	92.19	93.90	93.80	93.80
		0.4	ULSMV	82.74	91.19	85.22	87.44	92.40	85.19
		0.4	WLSMV	85.23	90.89	86.79	89.76	92.30	88.79
		0.4	BAYES	95.40	94.80	96.00	95.60	94.90	96.10
		0.7	ML	93.60	95.30	93.40	94.90	95.40	93.30
		0.7	ULSMV	91.50	94.70	90.20	93.30	94.40	91.20
		0.7	WLSMV	92.10	94.30	91.40	94.10	94.40	91.90
		0.7	BAYES	93.40	94.20	93.20	94.80	94.00	93.00
	2 factors (φ = 0.3)	0.4	ML	92.40	95.20	92.30	93.20	95.00	94.30
		0.4	ULSMV	87.49	93.50	87.33	88.93	93.00	90.58
		0.4	WLSMV	87.60	93.20	89.30	89.85	92.60	90.39
		0.4	BAYES	93.70	93.30	92.20	92.90	94.90	92.80
		0.7	ML	94.30	94.00	95.20	92.90	95.10	94.00
		0.7	ULSMV	92.70	94.30	94.40	94.60	94.00	94.60
		0.7	WLSMV	92.50	93.70	94.30	94.00	93.80	94.10
		0.7	BAYES	93.60	93.00	94.50	93.60	94.70	95.10
	2 factors (φ = 0.6)	0.4	ML	93.67	95.00	92.56	93.00	94.40	93.50
		0.4	ULSMV	94.07	94.29	92.41	92.08	93.40	91.86
		0.4	WLSMV	94.95	93.99	93.08	90.88	93.30	91.47
		0.4	BAYES	81.20	80.70	75.48	79.50	87.50	76.30
		0.7	ML	92.60	94.20	93.00	93.40	93.60	91.70
		0.7	ULSMV	94.10	94.70	93.80	94.80	93.80	93.70
		0.7	WLSMV	93.80	94.40	93.20	94.00	93.20	92.80
		0.7	BAYES	93.60	94.20	94.00	94.60	94.20	93.90
1000	2 factors (φ = 0)	0.4	ML	93.40	94.80	93.70	93.60	93.40	94.80
		0.4	ULSMV	90.17	93.60	90.08	91.30	92.80	91.70
		0.4	WLSMV	90.57	93.60	91.18	91.80	92.40	92.60
		0.4	BAYES	94.80	95.70	95.40	95.00	93.10	95.70
		0.7	ML	95.50	95.10	95.10	95.20	94.70	94.40
		0.7	ULSMV	94.40	94.80	93.50	92.90	93.80	94.10
		0.7	WLSMV	94.90	94.70	93.90	93.80	93.70	94.20
		0.7	BAYES	95.10	95.10	94.40	95.00	93.80	95.00
	2 factors (φ = 0.3)	0.4	ML	93.10	93.60	93.50	95.10	94.70	94.80
		0.4	ULSMV	91.78	92.50	90.70	93.30	93.80	92.90
		0.4	WLSMV	92.08	92.50	91.30	93.10	93.80	93.30
		0.4	BAYES	93.70	94.80	93.90	94.90	94.30	94.30
		0.7	ML	96.20	94.60	93.80	94.10	94.30	95.00
		0.7	ULSMV	96.00	94.10	93.30	94.90	93.60	94.40

Sample Size	Model	Mean Factor Loading	Estimation Method	Number of Items = 10			Number of Items = 20		
				Left-Skewed	Normal	Right-Skewed	Left-Skewed	Normal	Right-Skewed
	2 factors (φ = 0.6)	0.7	WLSMV	95.70	94.10	93.60	94.80	93.60	94.20
		0.7	BAYES	96.30	94.30	93.80	94.70	94.50	95.10
		0.4	ML	95.50	94.50	95.30	94.20	95.00	93.30
		0.4	ULSMV	94.78	94.00	94.58	93.30	95.00	93.10
		0.4	WLSMV	94.58	94.00	93.88	92.90	94.80	92.90
		0.4	BAYES	80.00	89.30	85.20	86.40	91.80	87.60
		0.7	ML	92.10	94.70	94.50	91.70	93.50	91.80
		0.7	ULSMV	94.30	95.40	95.80	94.50	94.50	95.50
		0.7	WLSMV	93.90	95.10	95.70	94.10	94.40	95.70
		0.7	BAYES	93.00	94.70	95.00	95.70	94.90	96.10

Appendix 7. Mean, maximum and minimum values of r-seb

Sample Size	Model	MFL	Method	Number of Items = 10									Number of Items = 20								
				Left-Skewed			Normal			Right-Skewed			Left-Skewed			Normal			Right-Skewed		
				Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
200	Unidimensional	0.4	ML	0.90	0.93	0.86	0.98	1.01	0.93	0.91	0.95	0.85	0.96	1.00	0.91	0.98	1.03	0.93	0.97	1.01	0.90
200		0.4	ULSMV	0.76	0.90	0.73	0.96	0.99	0.91	0.77	0.95	0.71	0.90	0.98	0.85	0.97	1.02	0.93	0.91	1.02	0.84
200		0.4	WLSMV	0.72	0.86	0.68	0.94	0.97	0.89	0.74	0.92	0.69	0.77	0.96	0.66	0.95	0.99	0.90	0.77	0.97	0.70
200		0.4	BAYES	0.94	1.19	0.88	1.02	1.24	0.93	0.96	1.23	0.88	1.00	1.35	0.94	1.00	1.06	0.90	1.01	1.37	0.93
200		0.7	ML	0.97	1.02	0.92	1.00	1.03	0.97	0.98	1.01	0.96	0.99	1.03	0.96	0.99	1.03	0.96	0.99	1.03	0.96
200		0.7	ULSMV	0.96	1.00	0.92	0.99	1.04	0.96	0.95	0.98	0.93	0.96	0.99	0.91	0.99	1.03	0.96	0.96	1.01	0.92
200		0.7	WLSMV	0.92	0.96	0.89	0.97	1.01	0.94	0.93	0.96	0.92	0.93	0.97	0.89	0.97	1.00	0.94	0.93	0.98	0.91
200		0.7	BAYES	0.96	1.03	0.89	0.98	1.03	0.94	0.98	1.03	0.95	0.98	1.07	0.93	0.97	1.00	0.93	0.98	1.10	0.91
200	2 factors (φ = 0)	0.4	ML	0.73	0.78	0.69	0.89	0.94	0.85	0.76	0.78	0.70	0.88	0.94	0.83	0.95	1.00	0.92	0.89	0.95	0.83
200		0.4	ULSMV	0.77	0.96	0.69	0.86	0.92	0.82	0.75	0.95	0.66	0.73	0.91	0.66	0.92	0.97	0.88	0.73	0.91	0.68
200		0.4	WLSMV	0.74	0.93	0.66	0.86	0.91	0.82	0.75	0.91	0.65	0.60	0.75	0.55	0.90	0.96	0.87	0.60	0.74	0.54
200		0.4	BAYES	1.09	2.27	0.79	1.09	1.80	0.91	1.08	2.10	0.80	1.05	2.06	0.86	1.02	1.54	0.93	1.03	2.12	0.86
200		0.7	ML	0.95	0.97	0.92	0.99	1.02	0.96	0.96	1.00	0.93	0.98	1.04	0.94	0.99	1.04	0.96	0.99	1.02	0.96
200		0.7	ULSMV	0.94	0.97	0.92	0.98	1.01	0.95	0.95	0.98	0.91	0.95	1.00	0.89	0.99	1.02	0.95	0.96	1.02	0.93
200		0.7	WLSMV	0.89	0.92	0.86	0.97	1.01	0.95	0.92	0.95	0.89	0.88	0.92	0.83	0.97	1.01	0.93	0.90	0.94	0.86
200		0.7	BAYES	0.99	1.21	0.92	0.97	1.09	0.91	1.01	1.18	0.90	1.00	1.14	0.94	0.97	1.04	0.94	1.00	1.21	0.93
200	2 factors (φ = 0.3)	0.4	ML	0.77	0.80	0.75	0.91	0.94	0.88	0.78	0.84	0.74	0.89	0.92	0.84	0.96	1.00	0.92	0.89	0.93	0.86
200		0.4	ULSMV	0.76	0.98	0.69	0.87	0.92	0.84	0.75	0.95	0.66	0.75	0.95	0.69	0.93	0.98	0.88	0.75	0.96	0.66
200		0.4	WLSMV	0.74	0.95	0.67	0.86	0.91	0.83	0.74	0.95	0.67	0.62	0.87	0.54	0.91	0.96	0.87	0.63	0.84	0.56
200		0.4	BAYES	1.09	2.14	0.80	1.04	1.63	0.86	1.06	2.08	0.76	1.02	2.07	0.87	1.02	1.54	0.92	1.02	1.97	0.86
200		0.7	ML	0.98	1.02	0.91	0.98	1.01	0.93	0.96	1.00	0.94	0.98	1.02	0.94	0.99	1.03	0.96	0.98	1.03	0.91
200		0.7	ULSMV	0.95	0.99	0.91	0.97	1.00	0.91	0.95	0.97	0.93	0.94	0.98	0.90	0.98	1.01	0.96	0.95	0.98	0.89
200		0.7	WLSMV	0.93	0.98	0.88	0.96	0.99	0.91	0.92	0.95	0.87	0.91	0.96	0.87	0.96	0.99	0.93	0.92	0.95	0.86
200		0.7	BAYES	1.01	1.29	0.91	0.95	1.05	0.88	0.99	1.19	0.91	0.99	1.13	0.91	0.97	1.03	0.94	1.00	1.16	0.93
200	2 factors (φ = 0.6)	0.4	ML	0.81	0.84	0.77	0.93	0.96	0.92	0.80	0.84	0.76	0.92	0.96	0.88	0.97	1.02	0.92	0.93	0.98	0.89
200		0.4	ULSMV	0.78	0.97	0.71	0.91	0.94	0.87	0.78	0.96	0.71	0.77	1.01	0.71	0.96	1.02	0.92	0.78	0.99	0.70
200		0.4	WLSMV	0.75	0.93	0.67	0.89	0.93	0.86	0.76	0.91	0.68	0.66	0.93	0.59	0.93	0.99	0.89	0.67	0.90	0.61
200		0.4	BAYES	1.04	1.87	0.79	1.04	1.48	0.92	1.02	1.88	0.77	1.03	1.86	0.89	1.00	1.44	0.90	1.04	1.85	0.88
200		0.7	ML	0.97	1.00	0.93	0.98	1.02	0.95	0.97	1.00	0.92	0.99	1.03	0.94	0.99	1.02	0.96	0.99	1.05	0.95
200		0.7	ULSMV	0.95	0.98	0.92	0.98	1.02	0.95	0.95	0.98	0.92	0.95	1.00	0.91	0.99	1.02	0.94	0.96	1.00	0.91
200		0.7	WLSMV	0.93	0.96	0.89	0.96	1.01	0.93	0.93	0.98	0.90	0.92	0.97	0.88	0.97	0.99	0.93	0.92	0.98	0.87
200		0.7	BAYES	0.99	1.24	0.92	0.95	1.02	0.91	0.99	1.17	0.92	1.00	1.18	0.95	0.98	1.03	0.94	1.00	1.23	0.94
500	Unidimensional	0.4	ML	0.97	1.01	0.94	1.00	1.03	0.97	0.98	1.01	0.94	0.99	1.02	0.95	0.99	1.02	0.96	0.98	1.03	0.96
500		0.4	ULSMV	0.92	0.95	0.89	0.99	1.02	0.97	0.94	0.97	0.90	0.96	0.99	0.93	0.98	1.01	0.95	0.96	1.00	0.93
500		0.4	WLSMV	0.94	0.97	0.91	0.98	1.02	0.96	0.94	0.97	0.90	0.95	0.99	0.92	0.97	1.00	0.94	0.95	0.99	0.92
500		0.4	BAYES	0.95	1.02	0.84	0.96	1.07	0.76	0.99	1.06	0.94	1.01	1.50	0.93	0.99	1.10	0.94	1.00	1.44	0.95
500		0.7	ML	1.00	1.06	0.97	1.00	1.04	0.97	0.99	1.03	0.97	1.00	1.03	0.97	1.00	1.03	0.96	1.01	1.06	0.96
500		0.7	ULSMV	0.99	1.04	0.96	1.00	1.04	0.97	0.98	1.01	0.95	0.98	1.00	0.94	1.00	1.03	0.97	0.99	1.04	0.95
500		0.7	WLSMV	0.98	1.03	0.95	0.99	1.03	0.96	0.97	1.00	0.94	0.97	0.99	0.94	0.99	1.02	0.96	0.98	1.03	0.94
500		0.7	BAYES	0.96	1.00	0.81	0.96	1.03	0.79	0.95	0.99	0.90	0.96	1.02	0.84	0.98	1.02	0.85	0.97	1.02	0.85
500	2 factors (φ = 0)	0.4	ML	0.92	0.94	0.90	0.97	1.01	0.95	0.93	1.01	0.89	0.97	1.04	0.93	0.98	1.04	0.94	0.97	1.00	0.94
500		0.4	ULSMV	0.85	0.91	0.82	0.94	0.97	0.92	0.87	0.92	0.82	0.92	0.96	0.87	0.97	1.03	0.93	0.92	0.95	0.87
500		0.4	WLSMV	0.87	0.92	0.84	0.93	0.97	0.92	0.88	0.94	0.84	0.93	1.00	0.89	0.96	1.02	0.92	0.92	0.95	0.86
500		0.4	BAYES	1.10	1.76	0.91	1.03	1.31	0.93	1.10	1.73	0.91	1.02	1.70	0.93	1.01	1.57	0.94	1.03	1.69	0.93
500		0.7	ML	1.00	1.03	0.97	1.00	1.08	0.95	0.99	1.03	0.96	1.00	1.05	0.95	1.00	1.05	0.96	1.00	1.05	0.96
500		0.7	ULSMV	0.98	1.02	0.95	0.99	1.08	0.94	0.98	1.02	0.95	0.98	1.03	0.93	0.99	1.04	0.95	0.98	1.02	0.93
500		0.7	WLSMV	0.98	1.02	0.95	0.99	1.08	0.94	0.98	1.02	0.95	0.97	1.03	0.92	0.98	1.03	0.94	0.97	1.01	0.93
500		0.7	BAYES	0.97	1.03	0.93	0.96	1.06	0.90	0.97	1.03	0.93	0.97	1.03	0.91	0.97	1.02	0.88	0.98	1.07	0.92
500	2 factors (φ = 0.3)	0.4	ML	0.94	0.97	0.91	0.98	1.00	0.94	0.97	1.02	0.95	0.97	1.01	0.93	0.99	1.03	0.94	0.98	1.02	0.94
500		0.4	ULSMV	0.87	0.91	0.83	0.95	0.97	0.92	0.89	0.93	0.86	0.93	0.98	0.90	0.99	1.03	0.94	0.94	0.98	0.90
500		0.4	WLSMV	0.88	0.91	0.84	0.94	0.97	0.91	0.90	0.94	0.87	0.93	0.98	0.89	0.98	1.01	0.93	0.94	0.97	0.90
500		0.4	BAYES	1.09	1.64	0.90	1.02	1.33	0.91	1.09	1.66	0.90	1.02	1.66	0.91	1.01	1.45	0.91	1.04	1.58	0.92
500		0.7	ML	0.98	1.01	0.95	1.01	1.05	0.98	0.98	1.00	0.96	1.01	1.06	0.97	1.01	1.05	0.96	0.99	1.03	0.95
500		0.7	ULSMV	0.97	1.00	0.95	1.00	1.03	0.96	0.98	1.00	0.95	0.98	1.03	0.92	1.00	1.04	0.95	0.97	1.00	0.95
500		0.7	WLSMV	0.97	0.99	0.94	0.99	1.04	0.96	0.97	0.99	0.94	0.97	1.03	0.92	0.99	1.03	0.95	0.96	0.99	0.93
500		0.7	BAYES	0.96	1.00	0.92	0.97	1.03	0.91	0.96	1.00	0.92	0.98	1.08	0.90	0.98	1.04	0.89	0.98	1.11	0.91
500	2 factors (φ = 0.6)	0.4	ML	0.96	0.98	0.94	0.99	1.01	0.97	0.95	1.00	0.92	0.98	1.04	0.95	0.99	1.03	0.96	0.98	1.01	0.93
500		0.4	ULSMV	0.90	0.95	0.85	0.97	0.99	0.94	0.89	0.93	0.84	0.93	1.00	0.89	0.98	1.02	0.95	0.95	0.98	0.90
500		0.4	WLSMV	0.89	0.94	0.85	0.96	0.98	0.93	0.89	0.93	0.84	0.93	1.00	0.88	0.97	1.01	0.94	0.94	0.97	0.89
500		0.4	BAYES	1.04	1.44	0.92	1.02	1.32	0.92	1.04	1.44	0.89	1.02	1.49	0.93	1.00	1.24	0.91	1.02	1.46	0.92
500		0.7	ML	0.99	1.03	0.96	1.01	1.05	0.98	0.99	1.04	0.95	1.02	1.08	0.97	1.01	1.06	0.99	1.01	1.05	0.98
500		0.7	ULSMV	0.98	1.03	0.96	1.01	1.04	0.98	0.98	1.02	0.94	0.99	1.05	0.95	1.01	1.05	0.97	0.98	1.02	0.94
500		0.7	WLSMV	0.97	1.03	0.94	1.00	1.03	0.97	0.97	1.02	0.93	0.98	1.04	0.94	1.00	1.04	0.96	0.97	1.01	0.94
500		0.7	BAYES	0.97	1.01	0.91	0.97	1.01	0.92	0.97	1.05	0.92	0.99	1.13	0.92	0.98	1.04	0.91	0.99	1.08	0.93
1000	Unidimensional	0.4	ML	0.97	1.01	0.94	0.99	1.04	0.96	0.99	1.03	0.95	1.01	1.04	0.98	1.00	1.03	0.95	0.99	1.03	0.97
1000		0.4	ULSMV	0.95	0.99	0.93	0.98	1.03	0.96	0.97	1.01	0.94	0.99	1.03	0.96	0.99	1.03	0.95	0.98	1.02	0.95
1000		0.4	WLSMV	0.95	0.98	0.92	0.98	1.03	0.95	0.97	1.01	0.94	0.99	1.02	0.96	0.99	1.02	0.95	0.97	1.01	0.94
1000		0.4	BAYES	0.93	0.97	0.74	0.95	1.04	0.71	0.95	1.00	0.82	1.00	1.08	0.95	0.97	1.05	0.83	0.97	1.02	0.93
1000		0.7	ML	1.00	1.02	0.98	1.01	1.04	0.95	1.00	1.02	0.97	1.01	1.06	0.95	1.00	1.04	0.97	1.01	1.05	0.97
1000		0.7	ULSMV	0.99	1.02	0.97	1.00	1.03	0.95	0.98	1.01	0.96	0.99	1.03	0.93	1.00	1.04	0.97	0.99	1.03	0.96
1000		0.7	WLSMV	0.99	1.01	0.96	1.00	1.03	0.94	0.97	1.00	0.96	0.98	1.02	0.92	1.00	1.04	0.97	0.99	1.03	0.95
1000		0.7	BAYES	0.95	1.00	0.78	0.96	1.01	0.80	0.94	1.03	0.80	0.98	1.03	0.86	0.97	1.06	0.78	0.98	1.05	0.90
1000	2 factors (φ	0.4	ML	0.98	1.02	0.94	0.98	1.00	0.93	0.97	1.00	0.93	0.99	1.01	0.96	1.01	1.05	0.97	0.99	1.06	0.96
1000		0.4	ULSMV	0.91	0.94	0.87	0.96	0.99	0.92	0.92	0.94	0.89	0.97	0.99	0.93	1.00	1.04	0.97	0.97	1.04	0.94
1000		0.4	WLSMV	0.92	0.95	0.88	0.96	0.99	0.91	0.92	0.95	0.89	0.96	0.99	0.93	1.00	1.04	0.96	0.97	1.03	0.94
1000		0.4	BAYES	1.04	1.26	0.94	1.00	1.13	0.92	1.03	1.37	0.93	0.98	1.21	0.92	1.01	1.21	0.95	1.00	1.31	0.92

Sample Size	Model	MFL	Method	Number of Items = 10									Number of Items = 20								
				Left-Skewed			Normal			Right-Skewed			Left-Skewed			Normal			Right-Skewed		
				Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
1000		0.7	ML	1.00	1.03	0.96	1.01	1.03	0.97	1.00	1.04	0.96	1.01	1.05	0.95	1.00	1.04	0.96	1.02	1.08	0.98
1000		0.7	ULSMV	1.00	1.03	0.96	1.00	1.02	0.97	0.99	1.04	0.95	1.00	1.04	0.93	1.00	1.04	0.95	1.00	1.04	0.96
1000		0.7	WLSMV	1.00	1.03	0.96	1.00	1.02	0.97	0.99	1.04	0.95	0.99	1.03	0.93	0.99	1.03	0.95	0.99	1.04	0.96
1000		0.7	BAYES	0.97	1.02	0.87	0.97	1.06	0.93	0.97	1.01	0.93	0.95	1.06	0.79	0.97	1.04	0.84	0.97	1.13	0.84
1000	2 factors (φ = 0.3)	0.4	ML	0.97	1.01	0.95	0.99	1.01	0.96	0.99	1.02	0.97	0.98	1.02	0.95	1.00	1.03	0.95	0.99	1.03	0.95
1000		0.4	ULSMV	0.93	0.96	0.90	0.97	1.00	0.95	0.94	0.96	0.92	0.96	0.99	0.93	0.99	1.02	0.96	0.97	1.01	0.93
1000		0.4	WLSMV	0.93	0.96	0.90	0.97	0.99	0.94	0.95	0.97	0.92	0.96	0.99	0.93	0.99	1.02	0.95	0.96	1.01	0.92
1000		0.4	BAYES	1.04	1.27	0.95	1.00	1.13	0.92	1.05	1.42	0.93	0.97	1.18	0.90	1.00	1.11	0.96	0.99	1.38	0.90
1000		0.7	ML	1.00	1.02	0.96	1.01	1.07	0.96	1.01	1.04	0.97	1.01	1.07	0.96	0.99	1.04	0.96	1.01	1.07	0.95
1000		0.7	ULSMV	0.99	1.02	0.96	1.00	1.06	0.95	1.00	1.03	0.97	0.99	1.04	0.94	0.99	1.03	0.95	0.99	1.05	0.94
1000		0.7	WLSMV	0.99	1.01	0.96	1.00	1.06	0.95	1.00	1.03	0.97	0.99	1.05	0.93	0.98	1.03	0.95	0.99	1.05	0.93
1000		0.7	BAYES	0.97	1.05	0.91	0.97	1.02	0.90	0.97	1.05	0.93	0.95	1.04	0.79	0.96	1.04	0.83	0.97	1.13	0.85
1000	2 factors (φ = 0.6)	0.4	ML	0.97	1.00	0.94	0.99	1.03	0.97	0.97	1.02	0.93	1.00	1.03	0.97	0.99	1.04	0.97	1.00	1.04	0.94
1000		0.4	ULSMV	0.94	0.97	0.91	0.98	1.02	0.96	0.94	0.98	0.90	0.98	1.02	0.95	0.99	1.03	0.96	0.98	1.02	0.94
1000		0.4	WLSMV	0.94	0.97	0.91	0.98	1.01	0.95	0.94	0.98	0.90	0.97	1.01	0.95	0.98	1.02	0.96	0.98	1.02	0.94
1000		0.4	BAYES	1.02	1.23	0.91	1.01	1.13	0.96	1.01	1.32	0.89	0.99	1.20	0.92	1.00	1.10	0.95	1.00	1.23	0.92
1000		0.7	ML	1.01	1.04	0.95	1.02	1.05	0.99	1.00	1.04	0.95	1.01	1.03	0.96	1.00	1.02	0.97	1.01	1.03	0.99
1000		0.7	ULSMV	1.00	1.03	0.93	1.02	1.05	0.98	0.99	1.03	0.95	1.00	1.03	0.95	1.00	1.03	0.97	0.99	1.03	0.95
1000		0.7	WLSMV	1.00	1.03	0.93	1.01	1.05	0.98	0.99	1.02	0.94	0.99	1.03	0.94	1.00	1.02	0.96	0.99	1.02	0.95
1000		0.7	BAYES	0.99	1.08	0.90	0.98	1.08	0.90	0.96	1.03	0.85	0.95	1.02	0.81	0.97	1.03	0.82	0.97	1.08	0.85

Appendix 8. r-seb values of interfactor correlations

Sample Size	Model	Mean Factor Loading	Estimation Method	Number of Items = 10			Number of Items = 20		
				Left-Skewed	Normal	Right-Skewed	Left-Skewed	Normal	Right-Skewed
200	2 factors (φ = 0)	0.4	ML	0.62	0.84	0.67	0.85	0.93	0.84
		0.4	ULSMV	0.58	0.70	0.59	0.53	0.79	0.51
		0.4	WLSMV	0.56	0.68	0.58	0.34	0.77	0.34
		0.4	BAYES	1.23	1.10	1.28	1.15	1.05	1.13
		0.7	ML	0.97	0.99	1.01	0.99	1.05	0.99
		0.7	ULSMV	0.85	0.96	0.87	0.85	0.98	0.86
		0.7	WLSMV	0.61	0.94	0.62	0.54	0.96	0.55
		0.7	BAYES	0.98	0.99	1.00	0.98	0.99	0.99
	2 factors (φ = 0.3)	0.4	ML	0.67	0.88	0.70	0.83	0.94	0.89
		0.4	ULSMV	0.59	0.77	0.61	0.53	0.83	0.56
		0.4	WLSMV	0.56	0.76	0.58	0.35	0.81	0.34
		0.4	BAYES	1.18	1.06	1.16	1.07	1.10	1.13
		0.7	ML	0.96	0.96	0.96	0.98	1.00	0.97
		0.7	ULSMV	0.86	0.94	0.87	0.90	0.99	0.88
		0.7	WLSMV	0.72	0.93	0.71	0.77	0.97	0.74
		0.7	BAYES	0.96	0.96	0.96	0.98	1.01	0.97
	2 factors (φ = 0.6)	0.4	ML	0.77	0.91	0.80	0.89	0.93	0.84
		0.4	ULSMV	0.67	0.95	0.67	0.53	0.92	0.56
		0.4	WLSMV	0.65	0.93	0.69	0.44	0.89	0.47
		0.4	BAYES	0.98	1.08	0.95	1.12	1.12	1.09
		0.7	ML	0.97	1.00	0.98	0.97	0.99	0.97
		0.7	ULSMV	0.95	1.00	0.96	0.92	0.98	0.93
		0.7	WLSMV	0.91	0.98	0.94	0.88	0.96	0.90
		0.7	BAYES	1.01	1.01	1.03	1.01	1.00	1.01
500	2 factors (φ = 0)	0.4	ML	0.85	0.95	0.90	0.97	0.97	0.94
		0.4	ULSMV	0.68	0.86	0.72	0.81	0.91	0.78
		0.4	WLSMV	0.70	0.86	0.75	0.83	0.90	0.77
		0.4	BAYES	1.06	1.01	1.12	1.06	1.01	1.04
		0.7	ML	0.96	1.00	0.96	0.99	1.05	0.94
		0.7	ULSMV	0.91	0.98	0.90	0.97	0.99	0.90
		0.7	WLSMV	0.92	0.97	0.90	0.96	0.98	0.90
		0.7	BAYES	0.95	0.97	0.94	0.99	0.99	0.93
	2 factors (φ = 0.3)	0.4	ML	0.91	0.99	0.94	0.95	0.98	0.96
		0.4	ULSMV	0.77	0.92	0.77	0.84	0.93	0.86
		0.4	WLSMV	0.78	0.92	0.80	0.86	0.92	0.86
		0.4	BAYES	1.13	1.07	1.07	1.06	1.02	1.05
		0.7	ML	0.99	0.96	1.02	0.99	1.01	1.03
		0.7	ULSMV	0.96	0.95	0.99	0.99	1.00	1.00
		0.7	WLSMV	0.96	0.95	0.98	0.98	0.99	0.99
		0.7	BAYES	0.98	0.93	1.00	0.98	0.99	1.01
	2 factors (φ = 0.6)	0.4	ML	1.00	0.98	0.95	0.91	1.00	0.95
		0.4	ULSMV	0.92	0.96	0.89	0.84	1.00	0.91
		0.4	WLSMV	0.92	0.95	0.90	0.83	0.99	0.90
		0.4	BAYES	1.13	1.11	1.08	1.10	1.13	1.11
		0.7	ML	0.99	1.00	0.99	1.01	1.01	1.00
		0.7	ULSMV	0.99	1.01	0.99	0.99	1.01	0.98
		0.7	WLSMV	0.98	1.00	0.97	0.98	1.00	0.96
		0.7	BAYES	0.99	0.97	0.99	1.01	1.02	0.99
1000	2 factors (φ = 0)	0.4	ML	0.94	0.99	0.97	0.96	0.95	0.98
		0.4	ULSMV	0.83	0.94	0.85	0.89	0.92	0.91
		0.4	WLSMV	0.84	0.94	0.87	0.90	0.91	0.91
		0.4	BAYES	1.02	1.03	1.08	1.02	0.95	1.04
		0.7	ML	1.03	0.99	0.98	1.01	1.03	0.98
		0.7	ULSMV	0.99	0.99	0.97	0.97	0.99	0.97
		0.7	WLSMV	0.99	0.98	0.97	0.97	0.98	0.97
		0.7	BAYES	1.02	0.99	0.98	1.02	0.98	0.98
	2 factors (φ = 0.3)	0.4	ML	0.96	0.97	0.95	1.00	1.00	0.99
		0.4	ULSMV	0.88	0.95	0.87	0.95	0.98	0.93
		0.4	WLSMV	0.89	0.94	0.88	0.95	0.97	0.93
		0.4	BAYES	1.05	1.03	1.04	1.06	1.00	1.04
		0.7	ML	1.06	0.98	1.00	1.00	1.01	1.03
		0.7	ULSMV	1.04	0.98	0.98	0.99	1.00	1.00

Sample Size	Model	Mean Factor Loading	Estimation Method	Number of Items = 10			Number of Items = 20		
				Left-Skewed	Normal	Right-Skewed	Left-Skewed	Normal	Right-Skewed
	2 factors (φ = 0.6)	0.7	WLSMV	1.04	0.97	0.98	0.98	1.00	0.99
		0.7	BAYES	1.05	0.98	1.00	1.00	1.00	1.03
		0.4	ML	1.00	0.97	1.03	0.99	1.02	0.96
		0.4	ULSMV	0.96	0.95	0.98	0.96	1.03	0.94
		0.4	WLSMV	0.96	0.95	0.98	0.96	1.03	0.94
		0.4	BAYES	1.08	1.05	1.14	1.10	1.08	1.08
		0.7	ML	0.96	1.01	1.02	1.02	1.01	1.03
		0.7	ULSMV	0.95	1.01	1.03	1.00	1.01	1.03
		0.7	WLSMV	0.95	1.00	1.02	0.99	1.01	1.02
		0.7	BAYES	0.95	1.00	1.01	1.02	1.01	1.04

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 488–496

https://doi.org/10.21449/ijate.788078

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Investigation of the effect of online education on eye health in Covid-19 pandemic

Huseyin Kaya [1,*]

[1]Ophtalmology Department, Faculty of Medicine, Pamukkale University, Denizli, Turkey.

ARTICLE HISTORY

Received: July 07, 2020
Revised: Aug. 17, 2020
Accepted: Aug. 30, 2020

KEYWORDS

Covid-19,

Eye fatigue,

Eye health,

Online education

Abstract: The aim of this research is to evaluate the effect of online education on eye health in Covid-19 pandemic and to present a new scale on this subject. For this purpose, 402 students (257 females, 145 males) with a mean age of 20.26 from different faculties of Pamukkale university were asked about eye health by e-mail between 8-13 July 2020. Also, eye fatigue questionnaire was applied to evaluate eye fatigue. Corrected item-total correlations and Cronbach Alpha internal consistency coefficient techniques were used for reliability analysis. In this study, online education eye health scale in Covid-19 pandemic was found to be positively correlated with eye fatigue questionnaire. According to the results of simple linear regression analysis conducted to determine the predictive value of the online education eye health scale in Covid -19 pandemic to eye fatigue, it was found that the online education eye health scale in covid-19 pandemic significantly predicted eye fatigue. Data analysis were conducted with SPSS 21.0 statistical package program in 0.01 significance level.

1. INTRODUCTION

The novel coronavirus originated from a seafood market place at Wuhan, China. The zoonotic resource of SARS-CoV-2 is unclear, but, previous analysis suggested bats as the main key reservoir (Lu et al., 2020). As yet, no hopeful clinical treatments or prevention methods have been developed against human coronaviruses. The main transmission ways of coronaviruses are direct or indirect human contact, and viral droplets (Yuan et al., 2006). These transmission pathways lead to the rapid spread of the disease. Therefore, social distance and hygiene are very important in preventing the spread of the disease.

Coronavirus family had caused outbreaks in the past for example severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) (Wang et al., 2013; Zhong et al., 2003). SARS CoV-2 is responsible for Covid-19 pandemic worldwide. Covid-19 had some common symptoms like sore throat, cough, and fever (Tian et al., 2020). While Covid-19 may be asymptomatic or mild in most patients, it may be severe in some patients, leading to renal failure, respiratory failure and multiple organ failure (Chen et al., 2020; Huang et al., 2020). While typical symptoms were seen at the beginning of the pandemic, then atypical symptoms such as muscle pain, loss of taste or smell, and headache started to appear (Huang et al., 2020; Lee, Min, Lee & Kim, 2020).

*CONTACT: Huseyin Kaya ✉ hsynkaya@gmail.com Ophtalmology Department, Faculty of Medicine, Pamukkale University, Denizli, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

Eye fatigue-asthenopia consists of subjective complaints that cause discomfort in the eye (Gowrisankaran, Nahar, Hayes & Sheedy, 2012). Asthenopia manifests itself with complaints such as eye discomfort, tearing, dryness, blurred vision, inability to focus, foreign body sensation (Neugebauer, Fricke & Russmann, 1992). This is an important condition that affects attention and academic performance. In our age, the use of digital devices is increasing, depending on the technological developments. In addition, this period of use is increasing in the new generation. As a result, the risk of eye strain increases especially in young people. Considering the previous literature, it has been stated that asthenopia may be associated with various psychosocial and environmental factors. Prolonged near work, increased cognitive load, using computer/screen can affect the eye fatigue complaints (Agarwal, Goel & Sharma, 2013; Ostrovsky, Ribak, Pereg & Gaton, 2012).

The prevelance of eye fatigue was observed by previous studies. Han et al., (2013) reported the prevelance of 57% in Chinese students (Han et al., 2013). In another study, the prevalence of asthenopia was found to be 53.3% in collage students. Also workload, time spent on computer per day, sexuality and time spent on handheld digital devices were found sinificantly related eye fatigue/astenopia in this study (Xu, Deng, Wang, Xiong & Xu, 2019).

All social layers in society have been seriously affected by the Covid-19 pandemic. Especially people over the age of 65 have been the most restricted socially in this process. On the other hand, the education and training activities of young people were interrupted during this period. During this period, young people also had to stay at home. At the same time, online education activities have increased in this process. Online education has replaced face-to-face education widely all over the world. In this process, students were left alone with the screen for long hours. While this situation shapes their social relations and behavior patterns, it also affects the eye health.

The Covid-19 pandemic is one of the most important social events of the last century worldwide. The pandemic, which first started in China, spread to the whole world in a very short time and has seriously affected our country. Since the first case in our country, serious measures have been taken and the spread rate of the Covid-19 pandemic has been tried to be reduced. Within the scope of these measures, schools were closed and online education-training activities continued. In our study, we aimed to measure the effect of online education on eye health of university students. In addition, we aimed to look at the consistency of the scale we developed with this survey by applying eye fatigue questionnaire.

2. METHOD

2.1. Study Group

Our study group consisted of 402 university students who receive education in different faculties of Pamukkale University during the 2019-2020 academic year. Participants of this study are students of Faculty of Education, Faculty of Arts and Sciences, Faculty of Engineering, Kale Vocational School, Tavas Vocational School and Faculty of Medicine. 257 (63.9%) female and 145 (36.1%) male students were included in this study. The mean age of the participants was 20.26 years.

2.2. Procedure

First the literature on the concept of eye health in Covid-19 pandemic was reviewed and the knowledge and theories related to this field were analysed. A pilot test was created by looking at the related literature. During the creation of the pilot test, it was asked to 5 field and measurement/evaluation experts to reflect the test to be measured. The pilot test was arranged and applicated to an appropriate sample. The pilot test application was carried out with 78 university students in order to check whether the items in the scale would be comprehensible to students. This application was carried out by the researcher via online and students' feedbacks were taken into consideration. Based on the analysis performed on students' feedbacks, five items were removed from the draft scale. This way, the scale with four items became ready for test application. The items were determined by item-factor analysis. And also to get evidence construct validity Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were carried out. Finally the online education eye health scale in Covid-19 pandemic was formed. The flow diagram of the study is shown in Figure 1.

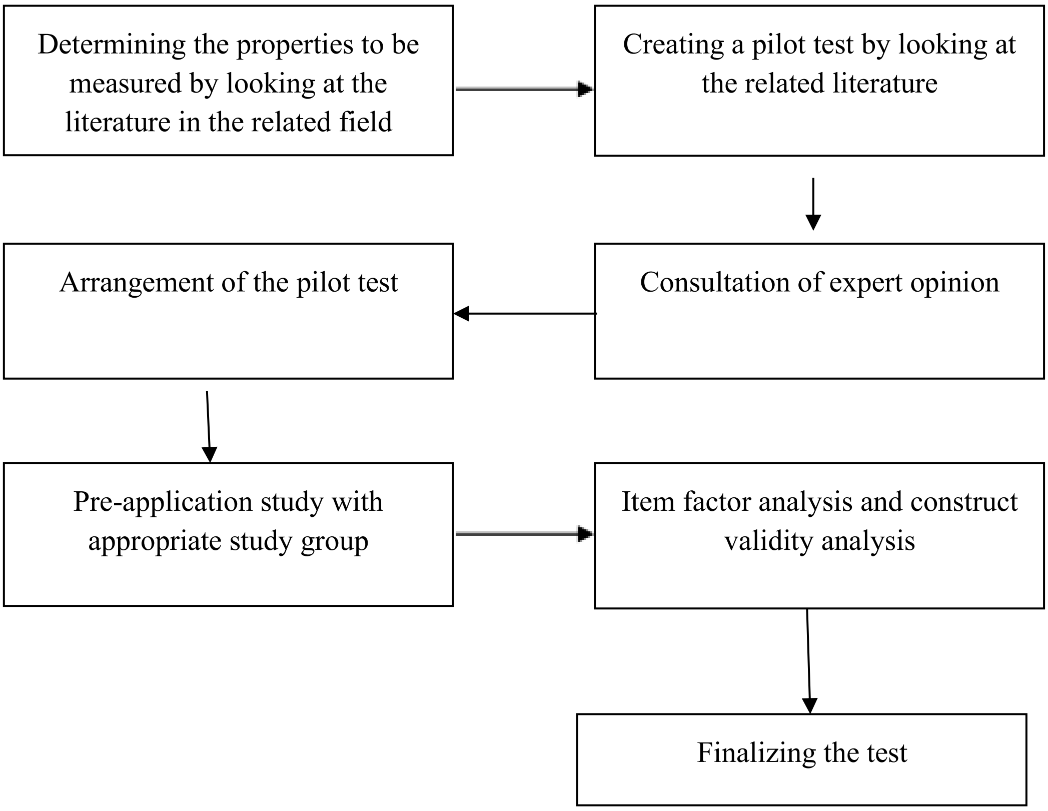

Determining the properties to be measured by looking at the literature in the related field
Creating a pilot test by looking at the related literature
Consultation of expert opinion
Arrangement of the pilot test
Pre-application study with appropriate study group
Item factor analysis and construct validity analysis
Finalizing the test

Figure 1. *The flow diagram of the study.*

The eye fatigue questionnaire consisted of 10 questions (tired eye, sore/aching eye, irritated eye, watery eye, dry eye, eye strain, hot/burning eye, blurred/doubled vision, difficulty in focusing/headache, visual discomfort). The online education eye health scale in Covid-19 pandemic was a four-item and one sub-dimensional scale The scale was a 3-point Likert type. The items of scale were 1: my eye health has not changed, 2: slight deterioration in my eye health 3: severe deterioration in my eye health. The eye fatigue questionnaire and online education eye health scale in Covid-19 pandemic were applied to university students by an e-mail. Before starting test necessary explanations were made. The tests were applied between 8-13 July 2020.

Statistical Analysis: Before starting statistical analysis, it was checked whether there was any missing data in the data set. After determining that the data set had a normal distribution (see Table 2 for skewness and kurtosis), the research data were analyzed. Cronbach Alpha technique was preferred for reliability analysis. Furthermore, Pearson correlation and simple linear regression analysis were used in the analysis of the data. The analysis was tested with the help of IBM SPSS program with a 0.01 level of significance.

3. RESULTS / FINDINGS

In this part of the study, construct validity analysis, reliability analysis, correlation and simple linear regression analysis are included.

3.1. Construct Validity

In order to determine properties of factorial design, Exploratory Factor Analysis (EFA), Before EFA, to test whether the sample size is sufficient for factoring, Kaise-Meyer-Olkin (KMO) test was carried out. As a result of analysis, KMO value was calculated to be .798. In accordance with this finding, sample size can be acknowledged to be "sufficient" for exploratory factor analyis (Field, 2009). Furthermore, results of Barlett's Test of Sphericty revealed that chi-square value was seen to be significant x^2= 922.98 (p<.001). After collecting these evidences about the suitability of the data set, factor analysis performed using the principal components analysis method (Tabachnick & Fidell, 2012). In the consequence of EFA, a single factor structure that explains 76.10% of total variance was obtained. In the result of the study, it was seen that item factor loads ranged from .79 to .83.

Table 1. *Finding Related to the Psychometric Properties (EFA and CFA) of Eye Health in the Covid-19 Period Scale*

Item No	EFA	CFA	
	λ^2	Standardized Coefficient	t-value (C.R)
1	.83	.95	4.91
2	.79	.75	10.29
3	.81	.93	6.02
4	.81	.77	10.22

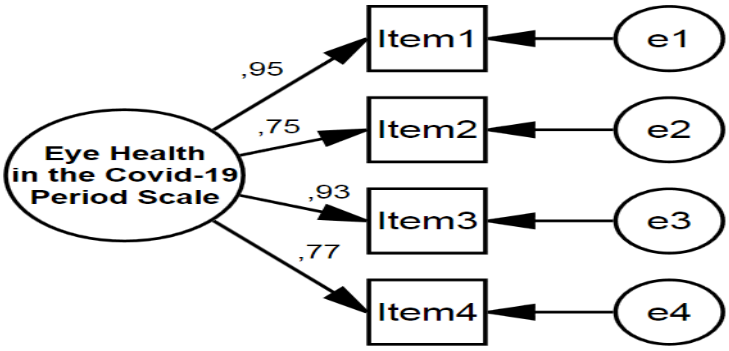


Figure 2. *Path Diagram and Factor Loadings of Eye Health in the Covid-19 Period Scale*

In the evaluation of the Confirmatory Factor Analysis (CFA), various fit indices are used. The frequently used ones are; chi-square fit (χ^2) and the ratio of chi-square to the degree of freedom (χ^2/sd), Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), Goodness of Fit Index (GFI), Adjustment Goodness of Fit Index (AGFI) and Standardized Root Mean Square Residual (SRMR) (Bayram, 2016; Brown, 2006). Less than 3 calculated χ^2/sd ratios, lower than .08 RMSEA and SRMR values, and bigger than .90 GFI, AGFI, and CFI values indicate the model data compatibility (Bayram, 2016; Brown, 2006; Schumacker & Lomax, 2010). The results of confirmatory factor analysis demonstrated that scale yielded a single factor (χ^2/sd= 0.22, p<0.001, RMSEA=0.00, SRMR= 0.00, GFI=1.00, AFGI=0.99, CFI=1.00). According to the obtained results, it can be stated that the Eye Health in the Covid-19 Period Scale possesses an acceptable level of model-data compatibility. In order to determine whether these values are acceptable, the C.R. (critical ratio) values, which are accepted as t-values in the AMOS program, were examined and each item was determined to be above the lower limit of 2.56 for significance at the .01 level. The t-values of the items on the scale ranged from 4.91 to 10.29. Accordingly, it can be stated that there is no need to remove any item from the scale and also the results of the confirmatory factor analysis indicated that the single factor structure fits well (Brown, 2006; Çokluk, Şekercioğlu, Büyüköztürk, 2014; Schumacker & Lomax, 2010).

Table 2. *Reliability Analysis of Eye Health Scale in Covid-19 Period*

Item No	Corrected item-total correlation	M (SD)	Skewness	Kurtosis
1	.84	2.05	-0.11	-1.71
2	.80	2.11	-0.23	-1.59
3	.80	2.08	-0.16	-1.62
4	.83	2.07	-0.15	-1.67

*Cronbach Alpha = 0.92

Corrected item-total correlations and Cranbach Alpha internal consistency coefficient analysis were used for the reliability of the online education eye health scale in Covid-19 pandemic. The adjusted item-total correlations of the scale have a value between 0.80 and 0.84. According to the analyzes, Cranbach Alpha reliability coefficient of the scale was obtained as 0.92 (Table 2).

Table 3. *Correlation Values Indicating Relationships Between Eye Health and Eye Fatigue in the Covid-19 Period*

	Eye Fatigue
Eye health in the Covid-19 period	.78**

***p*<0.01

According to the results of the analysis, a positive (r = .78, *p* <.01) correlation was found between eye fatigue and online education eye health scale in Covid-19 pandemic (Table 3).

Table 4. *Simple linear regression analysis results regarding the power of online education eye health scale to predict eye fatigue survey in Covid-19 period*

Predictor variable	R	R^2	F	B	Standart error B	*t*	*p*
Eye health in the Covid-19 period	.78	.62	652.44	4.20	0.16	25.54	.000

p<0.01

According to the simple linear regression analysis results, it was observed that the eye health scale significantly predicted eye fatigue in Covid-19 period. According to these analyzes, eye health in covid-19 period explained 62% of the total variance related to eye fatigue ($R^2 = .62$; $F_{Reg} = 652.44$; $p <.01$) (see Table 4).

4. DISCUSSION and CONCLUSION

The Covid-19 pandemic has profoundly affected all societies in the world, and has had many social, economic and psychological results. One of these results is the social isolation measures have been taken to slow the course of the disease. Schools and universities, where interpersonal distance cannot be maintained, are among the most easily spread environments. For this reason, it is very important to take necessary measures regarding education to reduce the speed of transmission of the epidemic (Afacan & Avcı, 2020). Accordingly, the Board of Higher Education has decided to close schools all over Turkey for three weeks from the date of March 16, 2020. Schools remained closed due to the continuing outbreak, and the Spring term was completed with online education. Although online education has the effect of reducing the transmission rate, it may have negative effects on eye health.

In this study, the effect of online education on eye health in Covid-19 period was investigated and a scale was developed on this subject. In addition, the relationship between eye health and eye fatigue in online education was investigated in Covid-19 period using scale. First of all, according to the analysis conducted for the scale, scale has been brought to the literature as a valid and reliable tool (see Tables 1, 2 and Figures 1, 2). With the developed scale, it was observed that the eye health of the university students was negatively affected by the online education of the Covid-19 pandemic process. In addition to this result, in the Covid-19 period, a positive correlation was found between the deterioration of eye health and eye fatigue in online education. In other words, eye fatigue increases as the result of online education deteriorate eye health.

In recent years, internet and screen usage has been increasing rapidly among the youth. Eye health can be negatively affected due to this increase. Previous studies have shown that eye health related to screen usage may be seriously affected. Digital screens like tablets, computers and mobile phones can cause harm by radiating short high energy waves that may penetrate eye tissues and can finally contribute to photochemical damage to the retinal cells. By this way, harmfull waves can cause a variety of eye problems ranging from dry eye to age-related macular degeneration (Bhattacharya, Saleem & Singh, 2020). It has been stated that as the duration of daily internet use increases, asthenopic complaints also increase significantly (Kaya, 2019). Another study indicated that computer use for more than 6 hours led to an increase in eye fatigue complaints (Agarwal, Goel & Sharma, 2013). In addition, it has been shown in previous studies that the symptoms of eye fatigue such as burning sensation, dryness, and tearing in the eyes due to the use of electronic devices such as computers and mobile phones have increased (Kaya, 2019; Kim, Lim, Gu & Park, 2017). In the study conducted by Kim et al. (2017), 59 participants used tablets and smart mobile device for 1 hour. Eye fatigue was evaluated before and 1 hour after using the tablet. According to this study, using tablets for 1 hour significantly increased the complaints of eye fatigue/asthenopia (Kim et al., 2017).

Environmental and social factors can also affect the eye health. In the study of Guo et al. on 1022 students; students' socioeconomic, dietary habits, lifestyles, eye-related symptoms, eye care habits and history of diseases were evaluated. In this study, it was investigated whether there is a relationship between fruit-vegetable consumption and the risk of asthenopia. According to the results of the study, it was found that dark-green leafy fruit consumption is associated with a lower risk of asthenopia (Guo et al., 2018). In the study conducted by Suh et al., (2018) on 60 patients, the patients slept in the laboratory for 3 nights. On the 3rd night, the patients slept in a 5-10 lux light environment. Eye fatigue findings were evaluated in the morning of the third day and on the fourth day. It was observed that eye strain, difficulty in focusing and ocular discomfort increased significantly in patients sleeping at 10 lux light intensity (Suh, Na, Ahn, & Oh, 2018).

4.1. Limitations and Suggestions

The study includes university students studying at various faculties of Pamukkale University. This can only give an idea about students studying at this university. Multicenter studies can give a wider idea about the subject. Also, trying to determine whether this new scale measures eye health in different age groups can be considered as a new research topic. This study is a quantitative research. In order to test the results of this study, a qualitative research on a similar subject may be proposed in the future**.**

In summary, it can be said that the validity and reliability of the eye health scale related to online education is sufficient in the Covid-19 period, which we prepared for the students who stayed at home during the Covid-19 period and thought that their eye health would deteriorate due to the use of more screen in addition to their normal use. In addition, it can be said that it was positively correlated with the eye fatigue questionnaire and its predictability was good.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. Permission was received from the Non-Interventional Clinical Ethics Committee of a University (dated 07.07.2020 and numbered 13). The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Hüseyin Kaya https://orcid.org/0000-0001-9633-3173

5. REFERENCES

Afacan, E., & Avcı, N. (2020). Koronavirüs (Covid-19) örneği üzerinden salgın hastalıklara sosyolojik bir bakış [A sociological overview of outstanding diseases through the coronavirus example]. *Eurasian Journal of Researches in Social and Economics, 7*(5), 1-14.

Agarwal, S., Goel, D., & Sharma, A. (2013). Evaluation of the factors which contribute to the ocular complaints in computer users. *Journal of Clinical and Diagnostic Research: JCDR*, *7*(2), 331–335. https://doi.org/10.7860/JCDR/2013/5150.2760

Bayram, N. (2016). *Yapısal eşitlik modellemesine giriş AMOS uygulamaları [Introduction to structural equation medeling SEM applications].* (3th ed.). Bursa: Ezgi Yayıncılık.

Bhattacharya, S., Saleem, S. M., & Singh, A. (2020). Digital eye strain in the era of COVID-19 pandemic: An emerging public health threat. *Indian Journal of Ophthalmology*, *68*(8), 1709–1710. https://doi.org/10.4103/ijo.IJO_1782_20

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York, NY: Guilford Press.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., … & Zhang, L. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet (London, England)*, *395*(10223), 507–513. https://doi.org/10.1016/S0140-6736(20)30211-7

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 practice]* (3th ed.). Ankara: Pegem Akademi Yayıncılık.

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

Gowrisankaran S, Nahar NK, Hayes JR & Sheedy JE (2012). Asthenopia and blink rate under visual and cognitive loads. *Optom Vis Sci* 89, 97-104

Guo, F., Zhang, Q., Fan, M. N., Ma, L., Chen, C., Liu, X. H., … & Liu, Y. (2018). Fruit and vegetable consumption and its relation to risk of asthenopia among Chinese college students. *International Journal of Ophthalmology*, *11*(6), 1020-1027. https://doi.org/10.18240/ijo.2018.06.21

Han C.C, Liu R, Liu R.R, Zhu Z.H, Yu R.B,, & Ma, L. (2013). Prevalence of asthenopia and its risk factors in Chinese college students. *Int J Ophthalmol*, *6*, 718–722

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., … & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet (London, England)*, *395*(10223), 497-506. https://doi.org/10.1016/S0140-6736(20)30183-5

Kaya H. (2019). Üniversite öğrencilerinde astenopik şikâyetlerin ve internet bağımlılığının ilişkisinin değerlendirilmesi [Evaluation of the relationship between asthenopic complaints and internet addiction in university students]. *Pamukkale Medical Journal 12*(3), 561-567.

Kim, D. J., Lim, C. Y., Gu, N., & Park, C. Y. (2017). Visual fatigue induced by viewing a tablet computer with a high-resolution display. *Korean Journal of Ophthalmology: KJO*, *31*(5), 388–393. https://doi.org/10.3341/kjo.2016.0095

Lee, Y., Min, P., Lee, S., & Kim, S. W. (2020). Prevalence and duration of acute loss of smell or taste in COVID-19 patients. *Journal of Korean Medical Science*, *35*(18), e174. https://doi.org/10.3346/jkms.2020.35.e174

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, …& Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet (London, England)*, *395*(10224), 565–574. https://doi.org/10.1016/S0140-6736(20)30251-8

Neugebauer A, Fricke J & Russmann W (1992). Asthenopia: frequency and objective findings. *Ger J Ophthalmol 1*, 122-124

Ostrovsky A, Ribak J, Pereg, A., & Gaton, D. (2012). Effects of job-related stress and burnout on asthenopia among high-tech workers. *Ergonomics 55*, 854-862

Schumacker, R.E., & Lomax, R.G. (2010). A beginner’s guide to structural equation modeling (3th ed.). New York, NY: Routledge.

Suh, Y. W., Na, K. H., Ahn, S. E., & Oh, J. (2018). Effect of ambient light exposure on ocular fatigue during sleep. *Journal of Korean Medical Science*, *33*(38), e248. https://doi.org/10.3346/jkms.2018.33.e248

Tabachnick, B.G., & Fidell, L.S. (2012). *Using multivariate statistics* (6th ed.). New York, NY: Harper Collins College Publishers.

Tian, S., Hu, N., Lou, J., Chen, K., Kang, X., Xiang, Z., ... & Zhang, J. (2020). Characteristics of COVID-19 infection in Beijing. *The Journal of infection*, *80*(4), 401–406. https://doi.org/10.1016/j.jinf.2020.02.018

Wang, N., Shi, X., Jiang, L., Zhang, S., Wang, D., Tong, P., … & Wang, X. (2013). Structure of MERS-CoV spike receptor-binding domain complexed with human receptor DPP4. *Cell research*, *23*(8), 986–993. https://doi.org/10.1038/cr.2013.92

Xu, Y., Deng, G., Wang, W., Xiong, S., & Xu, X. (2019). Correlation between handheld digital device use and asthenopia in Chinese college students: a Shanghai study. *Acta Ophthalmologica*, *97*(3), e442–e447. https://doi.org/10.1111/aos.13885

Yuan, J., Yun, H., Lan, W., Wang, W., Sullivan, S. G., Jia, S., … & Bittles, A. H. (2006). A climatologic investigation of the SARS-CoV outbreak in Beijing, China. *American journal of infection control*, *34*(4), 234–236. https://doi.org/10.1016/j.ajic.2005.12.006

Zhong, N. S., Zheng, B. J., Li, Y. M., Poon, Xie, Z. H., Chan, K. H., Li, P. H., … & Guan, Y. (2003). Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *Lancet (London, England)*, *362*(9393), 1353–1358. https://doi.org/10.1016/s0140-6736(03)14630-2

6. APPENDIX

If you would like to provide appendices, please provide here. You might put the scale items, if used in the study, or syntax, etc. if you wish to provide them.

Online education eye health scale in Covid-19 pandemic

1. In what way was your eye health affected in general, compared to the time before the epidemic, when you stayed at home and received online education due to the COVID-19 pandemic?

(1) My eye health has not changed

(2) Slight deterioration in my eye health

(3) Severe deterioration in my eye health

2. In what way did watching the lessons on your computer / tablet / mobile phone affect your eye health due to the COVID-19 pandemic?

(1) My eye health has not changed

(2) Slight deterioration in my eye health

(3) Severe deterioration in my eye health

3. How did doing homework on your computer / tablet / mobile phone affect your eye health during the COVID-19 pandemic?

(1) My eye health has not changed

(2) Slight deterioration in my eye health

(3) Severe deterioration in my eye health

4. During the COVID-19 pandemic, how did your use of more television / computer / mobile phones affect your eye health during the days you stayed at home?

(1) My eye health has not changed

(2) Slight deterioration in my eye health

(3) Severe deterioration in my eye health

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 497–521
https://doi.org/10.21449/ijate.727399
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **_Research Article_**

Prediction-Observation-Explanation (POE) Method and Its Efficiency in Teaching "Work, Energy, Power" Concepts

Turgay Nalkiran [1,*], **Sevilay Karamustafaoglu** [1]

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

ARTICLE HISTORY

Received: Apr. 27, 2020
Revised: July 21, 2020
Accepted: Aug. 29, 2020

KEYWORDS

Concept teaching,
POE,
Misconceptions

Abstract: With the developing technology educating students free from misconceptions, making sense of their learnings and using them in daily life are primarily aimed. This research is designed not only for teaching "work, energy and power" concepts and the relationships among them but also for investigating the effects of the teaching on students' achievements under the POE method. 6 students from the 9th grade studying at a private Anatolian High School chosen through easily accessible case sampling method, constituted the sample. 4 data collection tools (semi-structured interview, open-ended achievement test, concept map and concept cartoon) were applied. It was carried out within the scope of a single group pretest-posttest simple experimental study, a quantitative research method. For analyising the data, gap and content analysis methods were used. Thanks to the interviews, open-ended achievement test and concept map, as pre-tests, it was determined that students had many misconceptions about "work, energy, power" concepts and didn't have any scientific knowledge about the relationships among them. The students' drawings on these relationships were also far from scientific. After the concept teaching under POE was performed, the data collection tools were re-applied as post-tests. So, it was seen that students' misconceptions were largely eliminated by replacing them with scientifically-correct concepts and relationships as a result of that process. In the light of these findings, applying POE method in concept teaching on different classes, courses or subjects is higly recommended. Some suggestions are also made for the researchers wishing to work in this field.

1. INTRODUCTION

Waking up with many innovations and developments on every day, human has to keep up with them in order to survive. The importance of science is also increasing day by day. Not to be defeated and to be in secure in the economic and technological race that has been going on for centuries, make human follow scientific developments closely. Therefore, science education at schools has become extremely important in the relationship of the human with science. In "Science Curriculum", The Ministry of National Education MoNE (2018) in Turkey has aimed to gain many achievements that students can learn by doing and living and also which are important in terms of cognition, metacognition, sensory and psycho-motor skills. With the teaching of science lesson at schools, it is aimed to raise individuals who can think, research,

*CONTACT: Turgay NALKIRAN ✉ turgaynalkiran@gmail.com 🖃 Amasya University, Faculty of Education, Department of Educational Sciences, Amasya, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

inquire, discover, produce, use information without memorizing, and who are rational, scientific, open to communication and cooperation (Kaptan & Kuşakçı, 2002). However, it is stated that there are many factors that prevent students from reaching these goals in the education and training process, and also negatively affect their success. Misconceptions are foremost among these.

As one of the basic principles of science education, concepts must be structured in mind and associated with different schemes for students' processing information and using it in daily life without getting it ready. Because concepts are also at the heart of science lesson, the correct teaching of the them plays a key role in gaining the aims of science education. In the dictionary, "concept" term is defined as the general design of an object or thought in the mind (TLS, 2019). According to Ülgen (2001) concept is an information form showing the changeable common properties of objects and phenomena that people can visualize and make sense of in their minds. In the light of these definitions, "concept" may be explained as the common name of different kinds of objects which are capable of being transferred and grouped, also portrayed and interpreted in mind. Expressed as mental tools, concepts have a positive effect on people's thinking process and contribute to people not only for distinguishing one event, idea, thought, process from another but also for establishing a relationship by using them (Senemoğlu, 2013). The reason of this is that concepts having common features of objects, events and activities in number and species, have a certain relationship among themselves (Yel, 2015). Concepts, critical for learning in the educational process, reflect the characteristics of the events or objects that have made sense in the mind (Ülgen, 2001). In fact, concepts begin to be learned with the birth of a person. Until the end of human life, this process continuing from "easy and concrete concepts to complex and abstract concepts", will survive. In this period of time, some concepts are learned as a result of daily experiences or coincidences, while others are taught in a planned way in educational environments such as schools (Doğanay, 2005). Concept teaching one of the basic building blocks of the education and training process, is also considered as the first step in the realization of meaningful learning (Temizkan, 2011). In teaching of the concepts which have such a great importance and impact on learning process, several problems may occur.

Concept teaching includes theories such as social linguistics theory, social cognitive theory, constructivism (Baysen, Güneyli, & Baysen, 2012; Bozkurt, 2018; Hammer, 1996; Hein, 1991; Kocaman, 2006; Yağbasan, & Gülçiçek, 2003). In these theories, it is explained that generalization, distinction, definition, induction, deduction and both induction and deduction methods are used together in concept teaching. If concept teaching is not meaningful, some difficulties appear in effective learning.

For example some of concepts may be misleaded while some not being learned at all (Yılmaz & Çolak, 2011). The structures called misconceptions come first among these difficulties preventing learning of the concepts correctly (Byrd, McNeil, Chesney, & Matthews, 2015).

Concepts can be divided into two groups as abstract and concrete concepts. While concrete concepts can be perceived through the sense organs, abstract concepts cannot (Tokcan, 2015). There are many abstract concepts that students must acquire within the scope of science lesson during the education process at schools. If these concepts, which form the basis of science, cannot be understood and interpreted correctly, misconceptions that prevent establishing relationships between events and facts may arise (Ayyıldız & Altun, 2013). In this context, because science concepts have a more complex and abstract structure by its nature, more misconceptions can be seen in science education when compared other fields. The misconceptions defined in many ways in literature, appear when the concepts in human mind don't coincide with the scientifically-correct concepts (Nakhleh & Krajck, 1994). According to another definition, those are the problems which arise as a result of the inability in forming concepts correctly in a scientific way (Yağbasan & Gülçiçek, 2003). In terms of education, misconceptions are the kowledge incompatible with the scientific facts acquired by students before or during teaching process (Atılboz, 2004). Misconceptions come into view as a result of these misinformation, beliefs or experiences (Yenilmez & Yaşa, 2008). In other words, it is the mismatch between the concept definition created by students in their own minds and the scientifically-correct concept definition (Gönen & Akgün, 2005). In fact, because misconceptions are accepted as a major obstacle in learning the correct concepts, it is widely thought that not having any concepual knowledge is better than having misconceptions. On the other hand, concept teaching reconstructed with correct information, is also effective in eliminating the existing misconceptions (Ecevit & Şimşek, 2017). Because learners form concepts in an integrity in their minds, eliminating misconceptions that are inconsistent with current scientific information, is a difficult task. This integrity is also affected by students' daily experiences. So, these experiences may resist the positive change or development of the concepts. For this reason, possible misunderstandings of students may have a negative effect on their learning of the next concepts (Keçeli & Turanlı, 2013). The wrong concepts that students create in their minds also adversely effect the establishment of healthy connections with the new information or concepts. In other words, if the concepts in students minds are transferred to the learning stages after the existing misconceptions are eliminated, meaningful learning can be achieved (Atılboz, 2004). With the developing technology in the modern world, it is aimed to educate students not only wisely understanding what they learn and use in their daily lives, but also free from misconceptions. When the related literature is analyzed, it is seen that there are misconceptions about the concepts of energy, work, power (Avcı, & Karaca, 2012; Töman, & Çimer, 2016; Yürümezoğlu, Ayaz, & Çökelez, 2009).

Yürümezoğlu, Ayaz, & Çökelez, (2009) have found that middle school 6th, 7th and 8th grade students have deficiencies in structuring the concepts of energy, source of energy, form of energy and transfer of energy in their minds. Töman & Çimer, (2016) in their studies in which the misconceptions about the energy concept of students at different education levels are determined, have concluded that the misconceptions regarding energy issues and its concepts are continuing at every education level. Avcı & Karaca (2016) concluded that pre-service science teachers have misconceptions about the work concept because they cannot distinguish between daily work and physical work and they also confuse the work and power concepts.

Therefore, in order to eliminate misconceptions, the educational environments should be arranged in a manner appropriate for the implementation of new approaches where students' cognitively active participation can be achieved (Ayyıldız & Altun, 2013). As stated above, it is necessary to use a teaching method that is compatible with the features of the concepts which are aimed to be taught (Yel, 2015). In other words, it is important to apply appropriate teaching materials and activities within the framework of the teaching plan for preventing possible misconceptions. Because almost all students already have misconceptions, teaching in an environment where there is no misconception is a like dream for teachers (Koklu & Topcu, 2012). In the light of these data, in order to eliminate them, first of all, it is necessary to determine what causes to the misconceptions. Considering the related literature, many factors affect the students' misconceptions such as lack or insufficiency of prior knowledge, prior experiences and thoughts, the way teachers or textbooks are presented, insufficient concretization, lack of knowledge (Coştu, Ayas & Ünal, 2007). Due to the fact that students' daily lives and speeches are far from scientific, *the inability to interpret words, analogies and symbols correctly; insufficient pre-learnings; insufficient textbooks and materials in terms of content, shape and sampling; not using instructional strategies, methods and activities appropriate to the scientific development level of students* are also considered among the reasons for misconceptions (Aşçı, Özkan & Tekkaya, 2001).

Karaçam, & Gürsel, (2017) in their studies to determine how students mean "lifting force in liquids" in their minds, have found that the students copy the information about buoyancy as they are from sources such as textbooks and / or test books and have more stereotypical images with smooth geometric shapes. In this study, in order to correct the mental structures of the students towards lifting force in the direction of daily life based images, it has been proposed to take measures such as to include visuals based on daily life in materials like textbooks and test books and to organize training activities for teachers to deal with the issues of lifting force. Kurnaz, Tarakçı, Saydam, & Pektaş (2013) have examined the mental models of high school students related to electrification, lightning and lightning, and determined that they made non-scientific models. Researchers suggest using meaning analysis tables to reflect the differences between the three concepts and using conceptual change texts for possible misperceptions.

Many methods are used in both detecting and eliminating misconceptions that have various causes. The most frequently used ones are concept maps, concept networks, conceptual change texts, analogies, computer based learning methods (Atılğanlar, 2014). Apart from these, many methods such as information maps, concept puzzles, meaning analysis tables, word association tests, fishbone diagrams, structured grids, diagnostic branched trees, Vee diagrams, interview, drawings, multiple choice tests, educational games, open-ended success tests are also applied in both teaching concepts and eliminating misconceptions (Akyürek & Afacan, 2012; Başer & Çataloğlu, 2005; Çayan & Karslı, 2014; Tokcan, 2015).

One of the concept teaching process used not only to determine students' current prior knowledges and their scientific consistencies but also to eliminate misconceptions is Prediction-Observation-Explanation (POE) method (Tekin, 2006). As the name implies, POE is implemented in three stages as prediction, observation and explanation. Firstly, at prediction stage, students are requested to make predictions with their justifications regarding the possible outcome of the concept or event presented. By activating the pre-learnings in this way, misconceptions are detected by reaching their missing knowledge or wrong learnings, if any. Secondly, the observation stage which enables effective data collection on the relevant event or concept, is started. At this stage presentations, demonstrations or experiments are made about the event or concept presented to the students. Recording of the observations made before, during and after the experiment is also provided. Finally, at explanation stage the teacher explains the events or concepts according to the findings at the stage of prediction and observation. In other words, in which the lesson is taught, is started (Mpofu, 2006). The appropriate activities performed in the prediction, observation and explanation stages, also provide comprehensive information about students' concept structuring processes (Atasoy, 2002). This study was carried out on high school students and the effectiveness of the POE method was determined in terms of their learning of "work, energy and power" concepts. In this context, there is no study in the literature that is carried out with the POE method for the sample of this research and related concepts. Therefore, the results of the study are important for teaching these concepts. In addition, it is thought that it will be instructive physics and science teachers. This research will contribute to closing this gap in the literature by accompanying many studies on POE activities. It is also believed to be beneficial to scientists who will conduct research in this field.

This research is designed not only for teaching "work, energy and power" concepts and the relationships among them but also for investigating the effects of the teaching on students' achievements under the POE method. For this purpose, answers to the following research questions are sought:

1. What is the preliminary knowledge of the students about "work, energy and power" concepts at the beginning of the concept teaching process within the scope of POE?

2. What is the final knowledge of the students about "work, energy and power" concepts at the end of the concept teaching process within the scope of POE?
3. What are the effects of the POE method applied in teaching "work, energy and power" concepts?

2. METHOD

2.1. Research Model

In this study, the effectiveness of POE method in teaching 9th grade "work, energy, power" concepts was investigated. Unlike other studies, in this study interview about concepts, open-ended achievement test, concept map were used as measurement and evaluation tools in determining students' prior knowledge.

It was carried out within the scope of a single group pretest-posttest simple experimental study, a quantitative research method. This method was used because the 9th grade students in the school which was determined through an easily accessible sampling, had only one branch. In this context, the experimental group was created without the control group. In the cases where experimental and control groups can not be assigned randomly or there is no second group, it is stated that the application of single group research pattern does not constitute a problem for the validity of the research (Trochim, 2001). POE based materials properly developed for the research, were applied to the experimental group and its effects on the experimental group were investigated. In this way, it was aimed to observe the conceptual changes and developments of the students more clearly. In scientific researches, it is thought that the effects of simple experimental method applied on a single group will be high in observing the conceptual changes and developments in the participants (İpek Akbulut, Şahin & Çepni, 2013; Karslı & Çalık, 2012). In the framework of simple experimental method, *pre-tests* to determine students' prior knowledge, and *post-tests* to determine the achievement levels as a result of teaching, were applied. The aim of scientific researches is to examine the success development of the experimental group as a result of the concept teaching, however the obtained data can not be compared with a control group (Çepni, 2010).

2.2. Research Group (Participants)

The universe of this study aiming to teach "work, energy and power" concepts under POE method within the scope of "Energy" unit of 9th grade "Physics" course, contained all the high schools in Samsun province İlkadım district in 2019-2020 academic year. Among those, due to such factors as "easy access, being available at application time, having suitable conditions for the applicaitons" Anatolian High School was chosen through the easy sampling method, one of the purposeful sampling methods (Özmen & Karamustafaoğlu, 2019). Therefore, the study group of the research was formed with 6 students attending the 9th grade in this high school. Given the factors such as material, time and long efforts, purposeful sampling method can be used as the most appropriate method (Patton, 1990). Easily accessible sampling, which is a widely used method in scientific studies, is less costly than other methods. In addition, working with a recognized and known sample is effective in terms of bringing speed and feasibility to the research (Yıldırım & Şimşek, 2013). Taking into account the scientific research ethics, the names of the students participating in the study were not used. For this reason, the students were given such codes as *S*1*, S*2*, S*3*, S*4, *S*5 and *S*6 according to the interview order.

2.3. Process

After the sample was determined, the lesson plans were firstly prepared in order to express each stage of the concept teaching process under POE method in detail. Before the application, not only the achievement test consisting of open-ended questions but also the concept map developed by the researchers, were applied as pre-tests. In addition, a semi-structured interview was conducted with the aim of determining their memory elements such as episodes, images, propositions and indexes in order to comprehend how the participants put "work, energy and power" concepts into their minds. After reviewing relevant literature, data collection preparations for the interviews were completed and Anatolian High School was visited. The students, teachers and administrators of the school were informed about many subjects such as the purpose of the study, its importance, its contribution to science. Following the necessary approvals and permissions were obtained, the school was visited again and the school administrators were consulted about the place and time for the interviews. As a result of the desired answers, the semi-structured interview form was applied in an empty classroom, suitable for such factors as heat, light, silence, suitability for using, from 14.^{00} to 18.^{00} on 6th January 2020. The questions in the interview form were asked clearly (not having unceartain terms) when the participants felt themselves ready. During the interview the using of the gestures, mimics and words which might direct the participants, were especially avoided. Voice recording was also taken during the interviews to prevent data loss. In order to obtain further and more detailed data interviews, within the bounds of possibility were desired to be kept long, took an average of 30 minutes. After the interviews not only the students for their participation but also the teachers and the school administrators for their sincere support to the research were thanked. The findings obtained from the voice recordings were written down and also checked by an expert lecturer in this field. Therefore, it was aimed to increase the validity and reliability of the research. Researcher diversity may increase validity and reliability of scientific researches (Yıldırım & Şimşek, 2013). The day after the interviews, 7th January 2020, the open-ended achievement test and concept map were applied in the same class from 16.^{00} to 18.^{00}. Students were given 30 minutes for the open-ended achievement test and 3 minutes for the concept map as a response time. Then, open-ended achievement test and concept map documents were taken back and the participants were thanked again. The data obtained through the pre-test interviews, open-ended success tests and concept maps were examined in detail by the researchers. They also played an important role in the preparation of the lesson plans under POE method together with the findings obtained in the light of the literature review. The lesson plans prepared within the framework of POE method for teaching “work, energy, power” concepts and the relationships among them, were applied meticulously by the physics teacher (practitioner teacher) during the teaching period. As stated in the phsics curriculum, concept teaching under POE method in which different activities and practices were carried out at each stage, was carried out in 4 lessons (4 x 40 minutes). The first and second lessons were taught from 10.^{00} to 11.^{30} on 8th January 2020; third and fourth lesson were from 13.^{40} to 15.^{10} on 9th January 2020.

At the prediction stage: At the beginning of the first lesson, the concept cartoon previously developed by researchers, was firstly used. The concept cartoon papers, prepared for the relationships among “work, energy and power” concepts, named as *"Which Button Tells the Truth?",* were delivered to the students. Therefore, they were asked to find out which button was saying the truth with its explanation by giving 5 minutes as a response time as Figure 1. Thanks to this concept cartoon which highly attracted their attention, the students were both informed about the subject in the framework of the relevant curriculum and their motivation levels were also increased. At the same time, misconceptions of the students to the related concepts were determined.

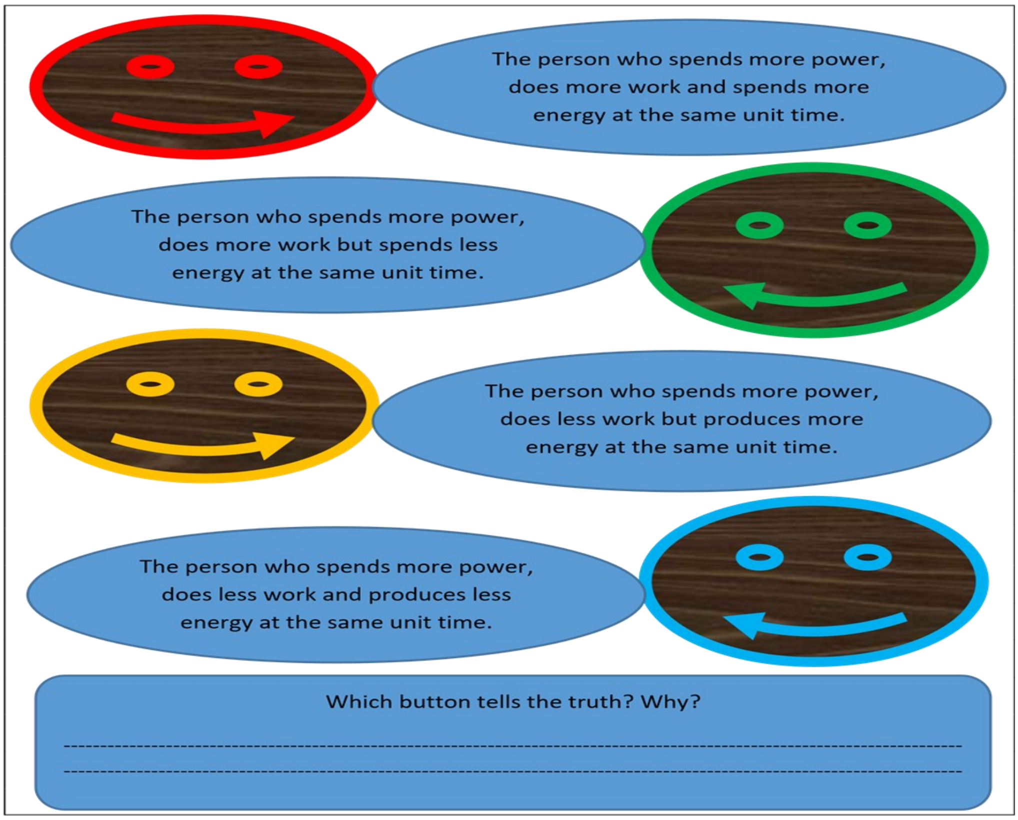

The person who spends more power, does more work and spends more energy at the same unit time.
The person who spends more power, does more work but spends less energy at the same unit time.
The person who spends more power, does less work but produces more energy at the same unit time.
The person who spends more power, does less work and produces less energy at the same unit time.
Which button tells the truth? Why?

Figure 1. *"Which Button Tells the Truth?" concept cartoon.*

At the observation stage: The animation video about "work, energy and power" concepts proposed by the Ministry of National Education, was watched at both the first and second lesson. The video paused when tought necessary, and some questions in the form of "What, How, Why, Which,…?" were directed. Therefore, the students' active participation to the observations and better understanding of the concepts and events were provided.

At the explanation stage: After evaluating the prediction and observation stages by comparing each other, instructional activities were carried out by emphasising on many issues such as definitions, uses and relationship of the relevant concepts. At this stage where the subject was taught, *the pendulum test* and *the slide experiment* were performed visually while *table-wall push experiment* was applied practically at second, third and fourth lessons. Through those experiments, the students' active participation in concept teaching process and realization of the subject were aimed. The next day, on 9th January 2020, the achievement test, concept map and concept cartoon used as pre-tests, were applied again in the same order as post-tests from $16.^{00}$ to $18.^{00}$. In addition, a week later than concept teaching process, on 15th January 2020, the semi-structured interview form was applied again from $14.^{00}$ to $18.^{00}$. A relevant transcript was also created. The data obtained as a result of those post-test applications were examined in detail by the researchers and expert opinion was again taken. The data obtained from both pre-test and post-test applications were analyzed with Mann Whitney U Test via an appropriate statistical program. Also, in the analysis of the interviews, appropriate programs headed for the qualitative data analysis were used and the findings were presented in tables.

2.4. Data Collection Tools

"Work, Energy and Power Achievement Test" consisting of open-ended questions, *"Work, Energy and Power Concept Map"*, *"Work, Energy and Power Concept Cartoon"* and *"Work-Energy and Power Semi-Structured Interview Form"* were used as data collection tools in this research where the effect of POE method on the teaching of "work, energy and power" concepts

were investigated. The reason for using more than one data collection tool in this study is to ensure triangulation and to determine how the relevant concepts are learned in depth.

In the light of the relevant literature review and the achievements in the "2018 Physics Teaching Program", expert opinion was also taken in creating of data collection tools and giving them their final shape. The whole concept teaching process under POE method and application of the data colloction tools were applied in January 2020.

2.4.1. Semi-Structured Interviews on "Work-Energy-Power" Concepts

The semi-structured interview form, which was used both as a pre-application pre-test and as a post-application post-test, consisted of 12 questions and extra questions directed according to the flow of the interview. In reference to the study's aims, the interview questions included many items related to memory elements (proposition, image, episode etc.) for "work, energy and power" concepts, such as detection, information, explanation and giving examples related to daily life. Also, through the 12th question of the interview, the participants were requested to draw the relationship of the concepts on a blank paper. By considering the achievements in the curriculum, the opinions of a physics teacher, a physics educator and a science educator were taken in the development of the interview questions related to the concepts.

An appropriate qualitative analysis program was used to transcribe, evaluate and analyze the data obtained from the interviews. Findings from the interviews are presented in Table 1, 2, 3 and 4.

2.4.2. "Work-Energy-Power" Open-Ended Achievement Test

The open-ended achievement test used both as a pre-application pre-test and as a post-application post-test, consisted of 6 questions. Participants' responses to open-ended achievement test questions were scored as "full comprehension 5 points", "partial comprehension 4 points", "no comprehension 3 points", "miscomprehension 2 points", "no response 1 point" (Abraham, Gryzybowski, Renner, & Marek 1992). The miscomprehension mentioned above refers to the misconceptions of the students. The findings are presented in Table 5.

2.4.3. "Work-Energy-Power" Concept Map

The concept map used both as a pre-application pre-test and as a post-application post-test, consisted of 2 concept boxes and 1 relationship box. 1 minute, as response time, was given for each of these gaps in the concept map designed according to the relationships among "work, energy and power" concepts. In scoring of the concept map, each concept and relationship box that was answered correctly was given "1" point. Appropriate computer programs were used to create, analyze, evaluate the data obtained from the concept map. The findings are presented in Tables 6, 7 and 8.

2.4.4. Work-Energy-Power Concept Cartoon

Work-Energy-Power Concept Cartoon" was developed to be used in the prediction phase of the activity developed for the POE method. In line with the opinions of a physics educator, a science educator, and a physics teacher, its validity was achieved and therefore applied.

"Which Button Tells The Truth?" consisted of 4 buttons in different colors (red, green, yellow, blue) prepared to show the relationships among "work, energy and power" concepts. Each button contained of different related to these relationships. In the concept cartoon where the "red button" tells the truth, "green, yellow and blue buttons" were located as distractors. The response gap was also reserved at the bottom of the cartoon for students to write down the reasons for the button they chose and 5 minutes was given as response time.

The concept cartoon was used in the prediction stage during the application process as a pre-test. After the concept teaching was completed, it was reapplied as a pos-test. Therefore, the relationships between the concepts were tried to be determined. Appropriate computer programs were used not only to create concept cartoons but also to analyze and evaluate the obtained data. The findings are presented in Table 9.

2.5. Data Analysis

In this research the obtained data were analyzed through both gap and content analysis methods (Yıldırım & Şimşek, 2013). The relevant findings were also shown in tables. The data obtained from the achievement test and the concept map were analyzed via an appropriate statistical package program. As the sample size was less than 30 participants, "Wilcoxon Signed Rank Test" one of the non-parametric tests, was used to determine whether there is a significant difference between the averages of the measurements made according to the pre-test and post-test results. When the assumptions of the parametric test are not met, “Wilcoxon Signed Rank Test” can be applied for multiple measurements for the relevant sample to determine if there is a significant difference between the averages (Büyüköztürk, 2012). Content analysis is used not only to gather data related to each other within certain concepts and themes but also to interpret them by organizing them in a way that the reader can understand more easily (Yıldırım & Şimşek, 2013).

3. FINDINGS

3.1. Findings Through Semi-Structured Interviews on “Work-Energy-Power” Concepts

Within the framework of the research aims, semi-structured interviews about “work, energy, power” concepts and the relationships among them, were applied to the participants both before and after the application. The students were requested to tell whether they had any experience related to these concepts and to declare them if any. They were also wanted to give examples for the relevant relationships. In that way, it was aimed to detect the misconceptions "already existing" and "after education", if any. These findings analyzed through content analysis method, are presented in Table 1, 2 and 3.

The participants were also wanted to draw relationships among the concepts on a blank paper. Thanks to those drawings, it was aimed to determine students’ images related to these concepts. Obtained findings analyzed via content analysis method, are presented in Table 4.

Table 1. *Content analysis results of the interviews.*

Statements	1. Interview						2. Interview					
	Interviewers						Interviewers					
	S1	S2	S3	S4	S5	S6	S1	S2	S3	S4	S5	S6
Work is doing a business.	*	*	*									
Work is making an effort.				*	*							
Work is a profession.						*						
Work is the movement of an object in the direction of applied force.							*	*	*	*	*	*
I played very well in the match, so I did a very good work (Work-related memory).				*								
People in the series I watch, do no work but gossipping (Work-related memory).	*											
I did no work other than playing on the phone last night (Work related memory).					*							
The work done in the animation we watch in the classroom (Work-related memory).								*		*	*	
Pushing the table and pushing the wall experiment (Work-related memory)							*		*			*

Energy is the power needed to do something.	*		*	*		*						
Energy is the force needed to do something.		*			*							
Energy is the capacity of something to do work							*	*	*	*	*	*
There are renewable and non-renewable energy types.	*	*	*	*		*						
There are nuclear energy types.					*							
There are types of energy as potential and kinetic.							*	*	*	*	*	*
Potential and kinetic energy can transform into one another.							*	*	*	*	*	*
They made wind energy tribunes to our district (Energy-related memory).			*									
We had solar panels installed in our house in the village last year (Energy-related memory).					*							
The pendulum experiment shown by the teacher (Energy types-related memory).							*		*			*
Sliding down the slide experiment (Energy-related memory).								*		*	*	
Power is the force applied to something.	*			*								
Power is the energy spent on doing a work.		*	*			*						
Power is the capacity to do a work.					*							
Power is the work done or energy spent per unit time.							*	*		*	*	*
Power is the energy spent to do a work.									*			
We won the match because we were more powerful (Power-related memory).				*								
Batuhan punched the door. His hand came from the opposite side because he is very powerful (Power-related moment).		*										
Last night I just played on the phone and slept, so I didn't waste any power (Power-related memory).						*						
Power experiment in animation watched in the classroom (Power-related moment).							*		*	*		*
Pushing the table and pushing the wall experiment (Power-related memory)								*			*	
Energy and power are spent to do a job.	*			*	*	*						
To do a job, energy is consumed but power does not have to be wasted.		*	*									
There is a top model car. When you put gasoline, its energy is filled and it moves by spending power. The car's movement is a work (Example for the relationship between work, energy and power)	*											
We spend energy to push the table. The powerful one pushes it faster. Pushing the table is a work (Example for the relationship between work, energy and power)					*							
Pushing the table and pushing the wall experiment (Example for the relationship between work, energy and power)							*		*			*
The animation watched and sliding down the slide experiment(Example for the relationship between work, energy and power)								*		*	*	

In the light of the findings shown in Table 1, it is seen that the participants had misconceptions about “work, energy and power” concepts and the relationships among them in the pre-test interviews. When looked at the post-test interviews, as a result of the teaching under POE, it is understood that misconceptions were largely eliminated. In addition, the images of the participants regarding these concepts are presented in Table 2, as both before and after application.

Table 2. *Participant images for "work, energy and power" concepts.*

Concept	Pre-Application Images	Post-Application Images
Work	Doing a business, making an effort, labouring somewhere, trying hard, profession.	The movement of an object in the direction of force.
Energy	Being energitic, the power used, the force used, chocolate.	The capacity of something to do work.
Power	Difficulty, trouble, force, being strong.	The work done or energy spent per unit time, the energy spent to do a work.

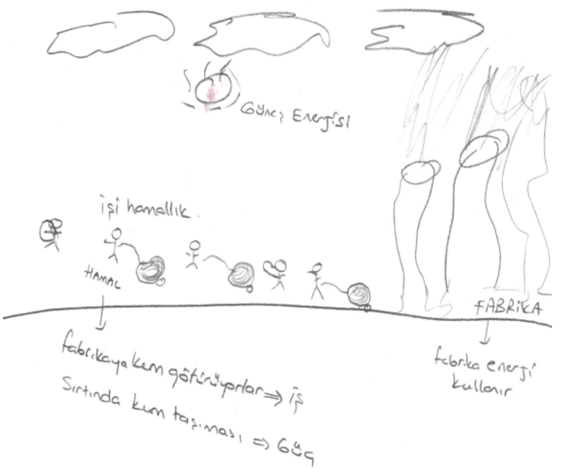
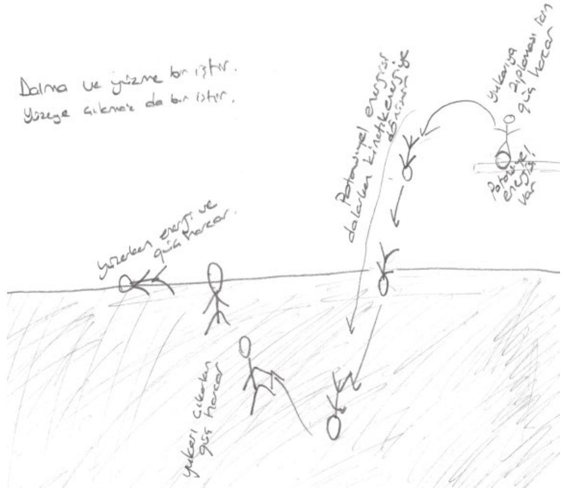
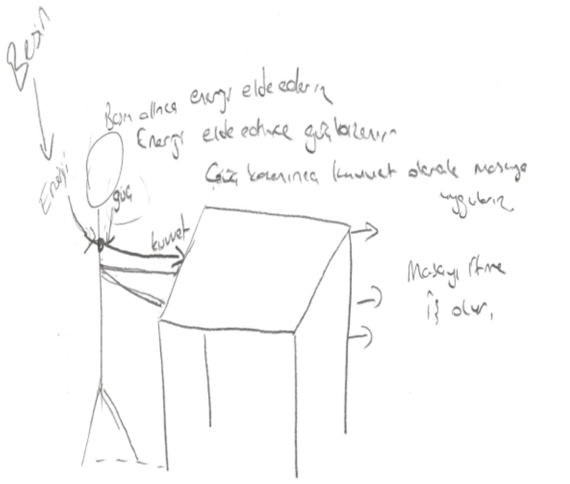
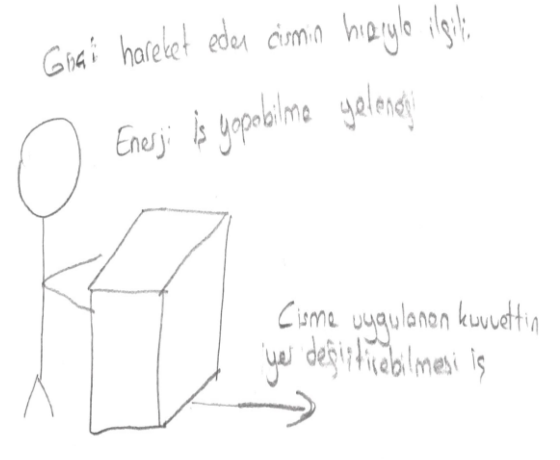
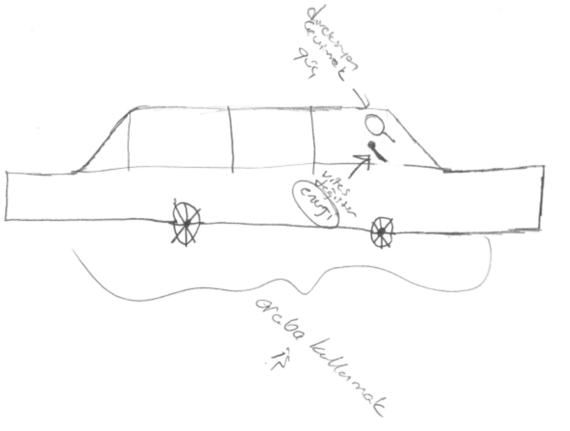
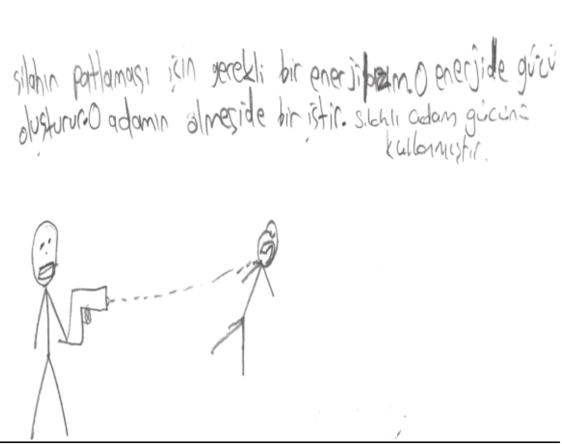
As seen in Table 2, it is understood that the students had misconceptions about "work, energy and power" concepts before the application. In the light of these data, the misconceptions were almost eliminated by turning them into scientific images. However, the statement of a participant (*S*3) on the concept of power as "*It is the energy spent to do a work*" revealed a new misconception. In addition, the episodes of the participants about these concepts are presented in Table 3, as both before and after the application.

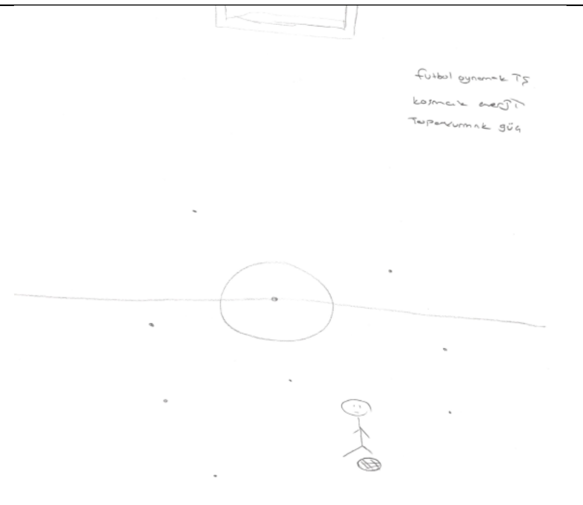
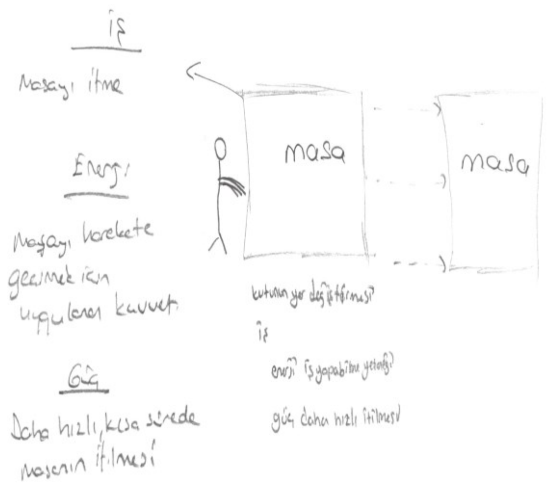
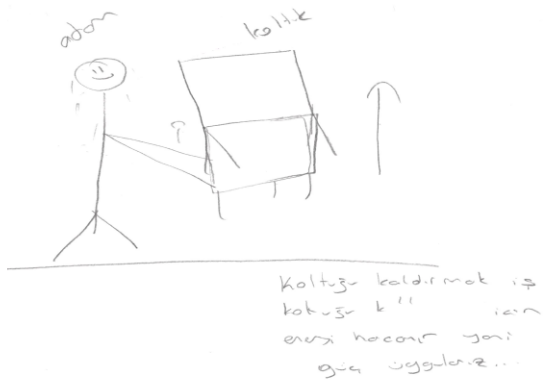
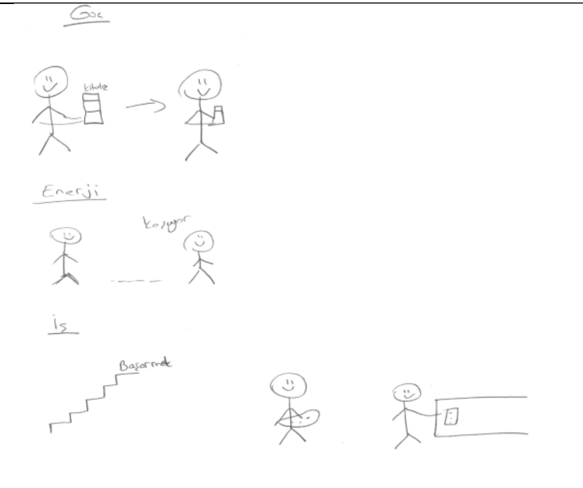
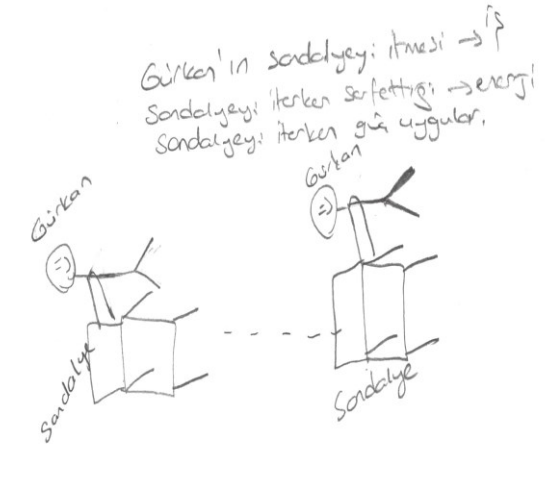
Table 3. *Participant episodes for "work, energy and power" concepts.*

Concept	Pre-Application Episodes	Post-Application Episodes
Work	I played very well in the match so I did a very good work, the people in the series I watch do no work but gossipping, I did no work other than playing on the phone last night, my father and I worked hard in the garden last Sunday, I work hard to be succesfull, I studied very hard to pass LGS exam last year.	The work done in the animation we watch in the classroom, pushing the table and pushing the wall experiment.
Energy	They made wind energy tribunes to our district, we had solar panels installed in our house in the village last year, I spent much energy to finish my homework yesterday, in a documantary I wacthed there a poisinous snake kills other beings via the huge energy in it, I was very tired yesterday and I didn't have a bit of energy.	The pendulum experiment shown by the teacher, sliding down the slide experiment.
Power	We won the match because we were more powerful, Batuhan punched the doo and his hand came from the opposite side because he is very powerful, last night I just played on the phone and slept, so I didn't waste any power, China lost its power because of Coronavirus, BMW is the best car because of its engine power, Ottoman Empire used its power to protect humanity.	Pushing the table and pushing the wall experiment, the animation watched and sliding down the slide experimen.

As stated in Table 3, according to the data obtained in the pre-test interviews, the participants had misconceptions about "work, energy and power" concepts. As a result of the concept teaching under POE, it is seen that the episodes related to these concepts before the application, were replaced by the episodes related to the experiments performed during the lesson and the activities in the relevant video. In addition, the drawings of the participants about the relationship of these concepts are presented in Table 4, as both before and after the application.

Table 4. *Participant drawings for the relationships among "work, energy and power" concepts.*

	Pre-Application Drawings	Post-Application Drawings
S1	 Güneş Enerjisi işi hamallık. HAMAL FABRİKA fabrikaya kum götürüyorlar ⇒ iş Sırtında kum taşıması ⇒ Güç fabrika enerji kullanır	 Dalma ve yüzme bir iştir. Yüzeye çıkmak da bir iştir. yüzerken enerji ve güç harcar. Potansiyel enerjisi dalarken kinetik enerjiye dönüşür yukarıya zıplaması için güç harcar Potansiyel enerjisi var yukarı çıkarken güç harcar
S2	 Besin Enerji güç kuvvet Besin alınca enerji elde ederiz Enerji elde edince güç kazanırız Güç kazanınca kuvvet olarak masaya uygularız Masayı itme iş olur,	 Güç: hareket eden cismin hızıyla ilgili. Enerji iş yapabilme yeteneği Cisme uygulanan kuvvetin yer değiştirebilmesi iş
S3	 direksiyon çevirmek güç vites değiştir enerji araba kullanmak iş	 silahın patlaması için gerekli bir enerji lazım. O enerjide gücü oluşturur. O adamın ölmeside bir iştir. Silahlı adam gücünü kullanmıştır.

S4		 İş Masayı itme Enerji Masayı harekete geçirmek için uygulanan kuvvet Güç Daha hızlı, kısa sürede masanın itilmesi masa masa kutunun yer değiştirmesi? iş enerji iş yapabilme yeteneği güç daha hızlı itilmesi
S5		 adam koltuk Koltuğu kaldırmak iş koltuğu k'' için enerji harcarır yani güç uygularız...
S6		 Gürkan'ın sandalyeyi itmesi → iş Sandalyeyi iterken sarfettiği → enerji Sandalyeyi iterken güç uygular. Gürkan Gürkan Sandalye Sandalye

As seen in Table 4, only one participant (*S*5) did not draw anything in the pre-test interviews. On the other hand, in the post-test interviews, all the students made drawings about the relationships among “work, energy and power” concepts.

According to the drawings in the pre-test interviews, it is seen that participants had misconceptions about the relationships of “work, energy and power” concepts. As a result of the concept teaching under POE, it is observed that the drawnings related to these relationships before the application, were replaced by thescientifically-correct drawings related to the experiments performed during the lesson and the activities in the relevant video.

3.2. Findings Through the Open-Ended Achievement Test

Frequency tables related to the participants' comprehension levels of the concepts are created according to the pre-test and post-test data and presented in Table 5.

Table 5. *Participants' comprehension levels frequencies related to the pre-test / post-test results.*

Questions	Comp. Levels Tests	N	FC f	PC f	NC f	MC f	NR f
What is work? Explain, please.	Pre-test	6	0	0	0	6	0
	Post-test	6	6	0	0	0	0
What is energy? Explain, please.	Pre-test	6	0	0	1	5	0
	Post-test	6	6	0	0	0	0
What are energy types? Explain, please.	Pre-test	6	0	5	0	1	0
	Post-test	6	5	1	0	0	0
What is power? Explain, please.	Pre-test	6	0	2	0	4	0
	Post-test	6	5	0	0	1	0
Is there a similarity or difference between work-energy-power concepts, if so, how is it? Explain, please.	Pre-test	6	0	1	0	5	0
	Post-test	6	3	3	0	0	0
Give examples for the relationship of work-energy-power concepts from daily life, please.	Pre-test	6	0	4	0	2	0
	Post-test	6	4	2	0	0	0
TOTAL	Pre-test	6	0	12	1	23	0
	Post-test	6	29	6	0	1	0

N: Sample number, *f:* Frequency, *FC:* Full comprehension, *PC:* Partial comprehension, *NC:* No comprehension, *MC:* miscomprehension, *NR*: No response (Abraham, Gryzybowski, Renner, & Marek 1992).

As can be seen in Table 5, there is an increase in post-test results compared to the pre-test results in all questions at full comprehension level. While the score of full comprehension level is 0 (zero) point in pre-application, it increases 29 x 5 = 145 points after application. This situation can be interpreted as a result of the success of the teaching in the elimination of the misconceptions. Although there isn't any change in the 1st and 2nd questions at the level of partial comprehension, a decrease in post-test results, from 12 to 6, is observed in the 3rd, 4th, 5th and 6th questions compared to the pretest results. While the pretest partial comprehension score was 12 x 4 = 48 points, the posttest score, declining 50%, was 6 x 4 = 24 points. This result can be interpreted as the concepts known as "partially" were largely learned "full" at the end of the application.

Merely in the 3rd question, only one student is at the level of no comprehension. 1 x 3 = 3 points obtained in the pre-test turned into 0 (zero) point in the post-test thanks to the elimination of the misconception after the application. In other questions, there aren't any students at "no comprehension" level both in the pre-test and post-test.

In all questions at the miscomprehension level, the frequency number was 23 and the score was 23 x 2 = 46 points before the application. This shows that students had many misconceptions within the frame of questions before the concept teaching under POE method was applied. As a result of the application, the frequency of the miscomprehensions decreased from 23 to 1 and the score also decreased to 2 points. These show that the concept teaching applied within the scope of POE was successful and so the misconceptions were largely eliminated. Since the students answered all questions, there isn't any participant at "no response" level in both the pre-test and post-test applications.

3.3. Findings Through the Concept Map

In the teaching process in order to eliminate misconceptions under POE method, the developed concept map was used as both a pre-test and a post-test. In the evaluation of concept maps, each concept and relationship box answered correctly, was given "1" point. The obtained results are presented in Table 6.

Table 6. *Concept map scoring table.*

Participants	PRE-TEST			POST-TEST		
	1. Concept Box "Potential energy"	2. Concept Box "Power"	Relationship Box "Because of the action of a substance"	1. Concept Box "Potential energy"	2. Concept Box "Power"	Relationship Box "Because of the action of a substance"
S1	0	0	0	1	1	0
S2	0	0	0	1	1	0
S3	0	0	0	1	1	1
S4	0	1	0	1	1	1
S5	0	0	0	1	1	1
S6	0	1	0	1	1	1
Total Points	0	2	0	6	6	4

As seen in Table 6, none of the 6 students could correctly answer the first concept box (potential energy) and the relationship box (because of the action of a substance) in the concept map used as pre-test. Except for two students (*S*4 and *S*6), the second concept box (power) was also replied incorrectly. Thanks to the concept teaching performed under POE method, both the first and second concept boxes were answered correctly by all the participants. Except two students (*S*1 and *S*2), they also answered the relationship box correctly. Considering the pre-test (2 points) and post-test (16 points) scores, it can be said that the concept teaching conducted within the scope of POE was successful and largely eliminated the misconceptions.

In order to determine whether there is a significant difference between pre-test and post-test results, "Wilcoxon Signed Rank Test" was conducted related to the achievement test and concept map scores. The test results are presented in Table 7.

Table 7. *The results of "Wilcoxon Signed Ranks Test" related to the pre-test and post-test scores of both academic achievement test and concept map.*

Test Type	Pre-test and Post-test Measurement	N	Rank Average	Rank Total	*z*	p^*	*r*
Achievement Test	Negative Ranks	0	0.00	0.00	-2.271	0.02	0.93
	Positive Ranks	6	3.50	21.00			
	No Difference	0					
Concept Map	Negative Ranks	0	0.00	0.00	-2.251	0.02	0.92
	Positive Ranks	6	3.50	21.00			
	No Difference	0					

$p<0.05$

When Table 7 is examined, a significant difference was found between the pre-test and post-test scores of the academic achievement test and concept map in favor of the post-test scores ($z = -2.271$, $p < 0.05$).

In addition, not only the pre-test and post-test mean but also standard deviations related to the open-ended achievement test and concept map were calculated. The relevant descriptive statistics are shown in Table 8.

Table 8. *The results of descriptive statistics related to the pre-test and post-test scores of both academic achievement test and concept map.*

Measurement	Test Type	N	$\overline{X}$	SD
Pre-test	Academic Achievement Test	6	16.17	2.23
	Concept Map	6	0.17	0.41
Post-test	Academic Achievement Test	6	27.83	2.48
	Concept Map	6	2.67	0.52

N: Participant Number

In the light of the findings in Table 8, in both pre-test and post-test applications, it is seen that there is a significant change in the academic achievement test consisting of open-ended questions, and concept map. This increase in favor of the post-test, shows that the concept teaching under POE was successful both in learning of the concepts and in eliminating the misconceptions.

3.4. Findings Through the Concept Cartoon

The concept cartoon with four buttons in different colors designed for the relationships among “work, energy and power” concepts, was used both as a pre-test and a post-test. The results obtained from the concept cartoon, "Which Button Tells th Truth?", are shown in Table 9.

Table 9. *The results of "Which Button Tells the Truth?" concept cartoon.*

Participants	PRE-TEST				POST-TEST			
	Red Button	Green Button	Yellow Button	Blue Button	Red Button	Green Button	Yellow Button	Blue Button
S1		*			*			
S2				*				*
S3	*				*			
S4		*			*			
S5		*				*		
S6	*				*			
Total	2	4	0	1	4	1	0	1

Red Button: The person who spends more power, does more work and spends more energy at the same unit time, Green Button: The person who spends more power, does more work but spends less energy at the same unit time, Yellow Button: The person who spends more power, does less work but produces more energy at the same unit time, Blue Button: The person who spends more power, does less work and produces less energy at the same unit time. “Red button tells the truth”

Looking at Table 9, only two students in the pre-test found the correct answer by marking the red button without explaining their reasons. On the other hand, 3 students answered as green button and one as blue incorrectly. No student chose the yellow button. Two students who answered correctly in the pre-test (*S*3 and *S*6) also chose the red button by explaining the correct reason in the post-test. Two students (*S*1 and *S*4) who chose the green button in the pre-test gave the correct answer by explaining the red button in the post-test. Thus, the number of those who answered correctly in the pre-test increased at 50% in the post-test. Two students (*S*2 and *S*5) who gave the wrong answer by choosing the blue and green button in the pretest did not find the correct answer in the posttest by giving the same answers.

Increasing the correct answers in the pre-test from 2 to 4 in the posttest, with an 50% increase, and 4 students explaining their responses' reasons correctly (that was zero in pre-test) show that the concept teaching performed within the scope of POE was effective in eliminating misconceptions.

4. DISCUSSION, CONCLUSION

This study is designed to investigate the effect of POE method on teaching "work, energy and power" concepts in "Energy" unit of 9th grade Physics course. 4 data collection tools (semi-structured interview form, open-ended achievement test, concept map, concept cartoon) which were applied both as pre-test and post-test, were used to determine and eliminate misconceptions. According to the findings obtained from these data collection tools, this section is presented in 4 subtitles.

4.1. Discussion and Conclusion Related to the Findings Through Semi-Structured Interview Form

The semi-structured interview form involved many items related to "work, energy and power concepts" such as the describing, explaining of the participants' memory elements (proposition, image, episode etc.) and associating them with daily life. It consisted of 12 questions and extra questions directed according to the flow of the interview. Moreover, through 12th question of the interview, the participants were asked to draw about the relationship of "work, energy and power" concepts on a blank paper.

In the interviews applied as a pre-test, analyzes were made within the framework of the images at Table 2, episodes at Table 3 and the drawings on the relationships among these concepts at Table 4. In this context; it is determined that they had many misconceptions about these concepts such as:

- Related to "work" concept, *"Work is a profession."* and *"I didn't do any work other than playing on the phone last night.",*
- Related to "energy" concept, *"Energy is the force needed to do something."* and *"Energy is the force needed to do something.",*
- Related to "power" concept, *"Power is the capacity to do a work."*and *"We won in the match because we were more powerful."*

The same interview form was reapplied after the concept teaching under POE. While there are no students (zero) who can correctly define the concepts of work energy and power in the pre-test, as a result of the teaching, all the students (six) have correctly defined these concepts in the post-test. In addition, it is concluded from their expressions and drawings regarding these concepts' relationships that "there is a scientific change in students' images and episodes", "there is a significant increase in their comprehension of the concepts" and "the existing misconceptions have substantially been eliminated".

Concept teaching with the POE method reveals the deficient or incorrect prior knowledge of the students. This situation may arise from the fact that POE is a method that enables the structuring of the concept in the mind and increases motivation and so can achieve meaningful learning (Bilen, 2009; Özdemir, 2011). The high desire and motivation of the students during the application stages of the POE method and thereby getting very quick and successful results in correcting the misconceptions support this information.

4.2. Discussion and Results Related to the Findings Through Open-Ended Achievement Test

In the open-ended academic achievement test prepared for “work, energy and power” concepts and the relationship of them, contained 6 questions as “What is work? Explain, please.”, “What is energy? Explain, please.”, “What are energy types? Explain, please.”, “What is power? Explain, please.”, “Is there a similarity or difference among work-energy-power concepts, if so, how is it? Explain, please.”, and “Give examples for work-energy-power concepts from daily life, please.” were directed to the participants.

When the statements of 6 participants towards those 6 questions are analyzed, it is seen that none of the answers given is at the level of “full comprehension”. In addition, 12 of the 36 responses in total are at the level of “partial comprehension” and only 1 is at “no comprehension” level. The remaining 23 answers at the level of “miscomprehension” proves students’ misconceptions before the application. In this context, when we look at the answers given by students to the level of misunderstanding it is seen that they had many misconceptions such as;

- Related to “work” concept, “Work is an event related to energy and movement.” and “Taking a glass and put it from one place to another is a work”,
- Related to “energy” concept, “Energy is what is spent to do a work.” and “For example, we spend energy while running”,
- Related to “energy types”, “There are renewable and non-renewable energy types.” and “To illustrate, chocolate is a type of energy that cannot be renewed because it ends when you eat it”,
- Related to “power” concept, "Power is the force applied to do a work." and "For instance, for lifting this table, power is necessary",
- Related to the relationship of “work, energy and power” concepts, “There is a relationship between them. Because the person who does a work both spends energy and applies power”,
- To exemplify the relationship of “work, energy and power” concepts from daily life, “For example, lifting a desk is a work which both requires energy and cannot be done without power”.

There aren’t any students at “no response” level in both pre-test and post-test applications. After the concept teaching conducted within the scope of POE, the same success test was applied again. Looking at the answers given, it is seen that while no answer was given at "full comprehension" level in the pre-test, 29 of the 36 answers in total were at "full comprehension" level in the post-test. In the post-test not only the answers at "partial comprehension" level decreased from 12 to 6 but also the answers at the level of "miscomprehension" decreased from 23 to 1.

These answers were scored as "full comprehension 5 points”, “partial comprehension 4 points”, “no comprehension 3 points”, “miscomprehension 2 points”, “no respond 1 point". In this context, the pre-test score of the test was calculated as 85 and the post-test score as 171.

When the statements of 6 participants for these 6 questions are analyzed, it is seen that the number of correct answers, which was 0 (zero) at "full comprehension" level in the pre-test, increased to 29 in the post-test. The answers at "partial comprehension" level decreased to 6. The aimed concepts were taught at the level of "full comprehension" at 80.55%. Therefore the rate of the “partial comprehension” level in post test was decreased from 33.66% to 16.66%. These results show that the concept teaching under POE was successful. In addition, the fact that post-test scores’ rising 171 from 85 proves this result.

On the other hand, the answers’ at "miscomprehension" level decreasing from 23 to 1 in the post-test shows that the current misconceptions regarding the concepts of work, energy and power were eliminated at 95.76%. These results show that the concept teaching was successful both in determining and eliminating misconceptions arising from students' incomplete or incorrect learning. It was also efficient in establishing close relations between the concepts.

Considering the academic success of the students in the literature review, it is seen that the applications carried out under POE method, have a more positive effect compared to the traditional teaching methods (Chew, 2008; Palmer, 1995; Özdemir, 2011). The fact that the scores in the achievement test used in this research, increased approximately twice in favor of the post-test and the fact that the misconceptions were eliminated at 95.76% support this information.

4.3. Discussion and Conclusion Related to the Findings Through Concept Map

In the concept map designed for this study, the students' were given 3 minutes as response time. In the case that the first concept box were replied as "Potential energy", the second concept box as "Power" and the relationship box as "Because of the action of a substance", 1 point was given to each concept box and relationship box. As a result of the answers given by 6 participants to these three boxes, the pre-test score of the concept map was calculated as 2 out of 18 full points, and the post-test score was 16 points.

In the concept map applied as a pre-test, two of the three boxes (the first concept box and the relationship box) could not be answered correctly by any student, while only two students (*S*4 and *S*6) could answer the second concept box correctly. In contrast, in the post-test, all students answered the first and second concept boxes correctly, while only two students (*S*1 and *S*2) answered the relationship box incorrectly. In line with these data, it is seen that the concept teaching within the scope of POE was successful and the existing misconceptions were eliminated at 88.88%. The post-test concept map test scores' increasing to 16 from 2 proves this result.

POE method makes the necessary environment suitable for students in realizing scientific process skills such as using knowledge, using mental skills in order to judge the problem and organizing the results achieved (Anagün & Yaşar, 2009). As it shows the importance of students' being related to daily life, this feature of POE is remarkable in terms of concept teaching. In addition, Ayvacı and Özbek (2015) draw attention to the importance of teaching features of science in terms of providing students with scientific thinking skills and creating a positive perspective towards science.

POE is a predictor for success because of its affective characteristics such as being enjoyable, fun, intriguing, motivational also increasing the desire to strive and act carefully (Mısır, 2009; Özyılmaz, 2008). In this study, it was observed that students were willing to eliminate existing misconceptions and to learn "work, energy and power" concepts during the concept teaching process conducted within the scope of POE. The fact that the scores in the concept map test applied in this research, increased eight times in favor of the post-test and the fact that the misconceptions were eliminated at 88.88% support those expressions in literature.

4.4. Discussion and Conclusion Related to the Findings Through Concept Cartoon

"Which Button Tells The Truth?" concept cartoon used both during and after the application, consists of 4 buttons in different colors (red, green, yellow, blue) prepared to show the relationships among "work, energy and power" concepts. The button in each color, consists of different sentences containing the relevant relationships. Also, the response gap is reserved at the bottom of the cartoon for students to write down the reasons for the button they chose. In the concept cartoon where the "red button" tells the truth, "green, yellow and blue buttons" are located as distractors. The statements in these buttons are presented below:

- *Red Button*: The person who spends more power, does more work and spends more energy at the same unit time,
- *Green Button*: The person who spends more power, does more work but spends less energy at the same unit time,
- *Yellow Button*: The person who spends more power, does less work but produces more energy at the same unit time,
- *Blue Button*: The person who spends more power, does less work and produces less energy at the same unit time.

Concept cartoon was used for the first time in the prediction stage at the beginning of the lesson. In pre-test, only two students wrote the red button which was the correct answer, but both of them failed to explain the reason. After the concept teaching under POE, two students (*S*4 and *S*6) who wrote green button in the pre-test changed their answer and chose the red button as the correct answer. Thus, a total of four students answered correctly and also explained the reason correctly. On the other hand, two students who chose blue button (*S*2) and green button (*S*5) in the pre-test gave the same answers in the post-test.

In this context, as a result of the concept teaching conducted under POE, although the misconceptions of the four students were eliminated, there was no change in the two students' answers. Considering the fact that there is an increase of fifty percent in the number of correct answerers according to pre-test and the explanation of the reasons for their answers correctly in post-test, the teaching can be considered successful.

Concept cartoons are visual tools that are prepared to view a scientific concept from a different wiewpoint in the form of discussion through their characters (Koch, 2010). When POE is used especially for the teaching of science concepts, it is important for students to question the nature of the concepts and to realize the changes in their own ideas. In this way, it increases learning and understanding of concepts (Kabapınar, Sapmaz & Bıkmaz, 2003; Köseoğlu, Tümay & Kavak, 2002, Liew, 2004). In the concept cartoon applied in this study, the number of those who chose the red button, where the relationships of "work, energy and power" is correctly expressed, increased from 2 to 4 in the post-test. Moreover the number of correct explanations of the reason for choosing the red button increased from 0 to 4 in the post-test. These results support the statements in literature.

On the other hand, the fact that students do not want to accept new information in some cases prevents correcting misconceptions (Torosluoğlu Çekiç, 2011). For this reason, replacing misconceptions with correct information is considered to be a very difficult task (Başer & Çataloğlu, 2005; Çaycı, 2007; Osborne & Freyberg, 1985; Özdemir, 2012). Two students who gave the wrong answer by choosing the green and blue buttons and could not explain the reasons in the pretest. They also chose the green and blue buttons in the posttest by not explaining the reasons correctly. This result which means that the existing misconceptions of the two students, unlike other four students, weren't eleminated is in parallel with these expressions in the literature.

Apart from these, observing after the prediction stage, was very effective on students' learning. During the observation stage, thanks to the demonstrations and experiments which increased both motivation and participation, the teaching process became conceptually rich. At explanation stage, not only the comparison of the predictions with the observations but also the active using of the predictions, information and findings, supported the conceptual meaning and learning while the concepts and their relationship were being taught.

As a result of the findings obtained through these tools used as pre-test and post-test, most of the misconceptions were eleminated by replacing them with the scientifically-correct concepts. In addition, it can be concluded that the concept teaching was successful in terms of memory elements such as image, episode, and in this context, students were also positively affected in terms of perception, attitude and behavior.

Declaration of Conflicting Interests and Ethics

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

ORCID

Turgay NALKIRAN https://orcid.org/0000-0002-1581-7174
Sevilay KARAMUSTAFAOĞLU https://orcid.org/0000-0002-2852-7061

6. REFERENCES

Abraham, M. R., Gryzybowski, E. B., Renner, J. W., & Marek, A. E (1992). Understanding and misunderstanding of eighth graders of five chemistry concepts found in textbooks. *Journal of Research in Science Teaching, 29*, 105-120. https://doi.org/10.1002/tea.3660290203

Akamca Özyılmaz G. (2008). *İlköğretimde analojiler, kavram karikatürleri ve tahmin-gözlem açıklama teknikleriyle desteklenmiş fen ve teknoloji eğitiminin öğrenme ürünlerine etkisi [The effects of science and technology education based on analogies, concept cartoons and predict-observe-explain techniques on learning outcomes]*. Doctoral Dissertation. Dokuz Eylül University, İzmir.

Akyürek, E., & Afacan, Ö. (2012). Kavram çarkı diyagramı kullanılarak 8. sınıf öğrencilerinin "Hücre Bölünmesi" ünitesindeki kavram yanılgılarının belirlenmesi [Determining the 8th grade students' misconceptions in the unit of "cell division" by using roundhouse diagramming]. *International Journal of Curriculum and Instructional Studies*, *2*(3), 47-58.

Anagün, Ş. S., & Yaşar, Ş. (2009). İlköğretim beşinci sınıf fen ve teknoloji dersinde bilimsel süreç becerilerinin geliştirilmesi [Developing scientific process skills at science and technology course in fifth grade students]. *Elementary Education Online, 3*(8), 843-865.

Aşçı, Z., Özkan, S., & Tekkaya, C. (2001). Students' misconceptions about respiration. *Education and Science, 26*(120), 29–36.

Atasoy, B. (2002). *Fen öğrenimi ve öğretimi*. Ankara: Gündüz Education and Publications.

Atılboz, N. G. (2004). Lise 1. sınıf öğrencilerinin mitoz ve mayoz bölünme konuları ile ilgili anlama düzeyleri ve kavram yanılgıları [9th grade students' understanding levels and misconceptions about mitosis and meiosis]. *Gazi University Journal of Gazi Educational Faculty, 24*(3), 147-157.

Atılğanlar, N. (2014). *Kavram karikatürlerinin ilköğretim yedinci sınıf öğrencilerinin basit elektrik devreleri konusundaki kavram yanılgıları üzerindeki etkisi [The impact of concept cartoons on seventh grade students' misconceptions about simple electric cırcuits].* Unpublished MA Dissertation, Hacettepe University Institute of Educational Sciences, Ankara.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view.* New York: Holt, Rinehart and Winston Inc.

Avcı, D. E., & Karaca, D. (2012). Fen bilgisi öğretmen adaylarının iş konusundaki kavram yanılgıları [Misconceptions of Science teacher candidates about work]. *Pamukkale University Journal of Education*, *31*, 27-39.

Ayvacı, H. Ş., & Özbek, D. (2015). Fen teknoloji toplum dersi kapsamında yapılan uygulamaların fen bilimleri öğretmen adaylarının bilimin doğası algılarına etkisi [The effect of science technology society course on preservice science teachers' perceptions of nature of science]. *HAYEF: Journal of Education, 12*(1), 93-108.

Ayyıldız, N., & Altun, S. (2013). Matematik dersine ilişkin kavram yanılgılarının giderilmesinde öğrenme günlüklerinin etkisinin incelenmesi [An investigation of the effect of learning logs on remedying students' misconceptions concerning mathematics lesson]. *Hacettepe University Journal of Education, 28*(2), 71-86.

Başer, M., & Çataloğlu, E. (2005). Kavram değişimi yöntemine dayalı öğretimin öğrencilerin ısı ve sıcaklık konusundaki yanlış kavramlarının giderilmesindeki etkisi [Effect of conceptual change oriented instruction on remediation of students' misconceptions related to heat and temperature concepts]. *Hacettepe University Journal of Education, 29*, 43-52.

Baysen, E., Güneyli, A., & Baysen, F. (2012). Kavram öğrenme-öğretme ve kavram yanılgıları: Fen bilgisi ve Türkçe öğretimi örneği [Teaching & learning concepts and misconceptions: Science and Turkish teaching cases]. *International Journal of New Trends in Arts, Sports & Science Education (IJTASE)*, *1*(2), 108-117.

Bilen, K. (2009).*"Tahmin Et-Gözle-Açıkla" (TGA) stratejisine dayalı laboratuvar yaklaşımı ile hazırlanan etkinliklerin, fen bilgisi öğretmen adaylarının kavramsal başarılarına, bilimsel süreç becerilerinin gelişimine, biyoloji laboratuvarına yönelik tutumlarına ve bilimin doğasını hakkındaki görüşlerine etkisi [Predict-Observation-Explain" (POE) strategy compared to a verification laboratory approach on the development of pre-service science teachers' science skill processes and their views of nature of sceince in a general biology laboratory course].* PhD Dissertation, Gazi University, Ankara.

Bozkurt, B. Ü. (2018). Kavram, kavramsallaştırma yaklaşımları ve kavram öğretimi modelleri: Kuramsal bir derleme ve sözcük öğretimi açısından bir değerlendirme [Concepts, conceptualization approaches, and concept teaching models: A theoretical review and an evaluation in terms of teaching vocabulary]. *Language Journal, 169*(2), 5-24.

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

Byrd, C. E., McNeil, N. M., Chesney, D. L., & Matthews, P.G. (2015). A specific misconception of the equal sign acts as a barrier to children's learning of early algebra. *Learning and Individual Differences, 38*, 61-67. https://doi.org/10.1016/j.lindif.2015.01.001

Çepni, S. (2010). *Araştırma ve proje çalışmalarına giriş* (5. ed). Trabzon: Erol Offset.

Chew, C. (2008). *Effects of biology-infused demonstrations on achievement and attitudes in junior college physics.* Unpublished doctoral dissertation. The University of Western Australian, Australia.

Cinici, A., & Demir, Y. (2013). Teaching through cooperative POE tasks: A path to conceptual change. *The Clearing House: A Journal of Educational Strategiies, Issues and Ideas, 86*(1), 1-10. https://doi.org/10.1080/00098655.2012.712557

Coştu, B., Ayas, A., & Ünal, S. (2007). Kavram yanılgıları ve olası nedenleri: Kaynama kavramı [Misconceptions about boiling and their possible reasons]. *Kastamonu Educational Journal, 15*(1), 123-136.

Çayan, Y., & Karslı, F. (2014). 6. sınıf öğrencilerinin fiziksel ve kimyasal değişim konusundaki kavram yanılgılarının giderilmesinde probleme dayalı öğrenme yaklaşımının etkisi [The effects of the problem based teaching learning approach to overcome students' misconceptions on physical and chemical change]. *Kastamonu University Kastamonu Educational Journal, 23*(4), 1437-1452.

Çaycı, B. (2007). Kavram değiştirme metinlerinin kavram öğrenimi üzerindeki etkisi [The effect of conceptual change texts on the concept learning]. *Gazi Üniversitesy Journal of Gazi Educational Faculty, 27*(1), 87-102.

Doğanay, A. (2005). *Hayat bilgisi ve sosyal bilgiler öğretimi*. Ankara: Pegem Academy.

Ecevit, T., & Şimşek, P. Ö. (2017). Öğretmenlerin fen kavram öğretimleri, kavram yanılgılarını saptama ve giderme çalışmalarının değerlendirilmesi [The evaluation of teachers' science concept teaching and their action to diagnose and eliminate misconceptions]. *Elementary Education Online, 16*(1), 129-150. https://doi.org/10.17051/io.2017.47449

Gömleksiz, M. N. (2018). "Öğretim İlkeleri ve Yöntem Seçimi". Gömleksiz, M. N. (Ed.). *Öğretim İlke ve Yöntemleri* (p. 73-100). Elazığ: Asos Publications.

Gönen, S., & Akgün, A. (2005). Isı ve sıcaklık kavramları arasındaki ilişki ile ilgili olarak geliştirilen çalışma yaprağının uygulanabilirliğinin incelenmesi [The investigation of applicability of worksheet was developed about relationship between heat and temperature concepts]. *Electronic Journal of Social Sciences, 3*(11), 92- 106.

Hammer, D. (1996). How many alternative perspectives of cognitive structure influence instructional perceptions and intentions? *Journal of Learning Sciences. 5*(2), 97-127.

Hein, G.E. (1991). Constructivist learning theory, the museum and the needs of people. *CECA (International Committee of Museum Educators), Conference Jerusalem Israel, Leseley College*. Massachusetts, USA.

İpek Akbulut, H., Şahin, Ç., & Çepni, S. (2013). İş ve enerji konusu ile ilgili kavramsal değişimin incelenmesi: İkili yerleşik öğrenme modeli örneği [Examining conceptual change in work and energy topic: Dual situated learning model sample]. *Mehmet Akif Ersoy University Journal of Education Faculty. 13*(25), 241-268.

Kocaman, A. (2006). *Dilbilim: Temel kavramlar, dilbilim, temel kavramlar, sorunlar, tartışmalar.* Ed. A. Kocaman, Ankara: Language Association.

Koch, J. (2010). *Science stories science methods for elementary and middle school teachers* (4th edition). Canada: Cengage Learning.

Kabapınar, F. M., Sapmaz, N. A., & Bıkmaz, F. H. (2003). *Aktif öğrenme ve öğretme yöntemleri, fen bilgisi öğretimi.* Ankara: Ankara University Faculty of Educational Sciences Education Research and Application Center Publications.

Karslı, F., & Çalık, M. (2012). Can freshman science student teachers' alternative conceptions of 'elektrochemical cells' be fully diminished? *Asian Journal of Chemistry, 24*(2), 485-491.

Kaptan, F., & Kuşakcı, F. (2002). Fen öğretiminde beyin fırtınası tekniğinin öğrenci yaratıcılığına etkisi [The effect of brain storming technique on student creativity in science teaching]. *V. National Science and Mathematics Education Congress Proceedings Book* (p. 197-202). METU: Ankara.

Karaçam, S., & Gürsel, Ü. (2017). Lise öğrencilerinin sıvılarda kaldırma kuvveti kavramına yönelik görsel imgeleri ve imgenin kökenleri [High school students' visual ımages about the concept of buoyancy and roots of those images]. *Mehmet Akif Ersoy University Journal of Education Faculty,* 41, 326-345. https:// doi.org/10.21764/efd/14301

Kearney, M. (2004). Classroom use of multimedia-supported predict–observe–explain tasks in a social constructivist learning environment. *Research in Science Education, 34*, 427–453. https://doi.org/10.1007/s11165-004-8795-y

Kearney, M., Treagust, D. F., Yeo, S., & Zadnik, M. (2001). Student and teacher perceptions of the use of multimedia supported predict-observe-explain tasks to probe understanding. *Reserach in Science Education, 31*(4), 589-615. https://doi.org/10.1023/A:1013106209449

Keçeli, V., & Turanlı, N. (2013). Karmaşık Sayılar konusundaki kavram yanılgıları ve ortak hatalar [Misconceptions and common errors in complex numbers]. *Hacettepe Üniversity Journal of Education, 28*(1), 223-234.

Koklu, O., & Topcu, A. (2012). Effect of Cabri-assisted instruction on secondary school students' misconceptions about graphs of quadratic functions. *International Journal of*

Mathematical Education in Science and Technology, 43(8), 999-1011. https://doi.org/10.1080/0020739X.2012.678892

Köseoğlu, F., Tümay, H., & Kavak, N. (2002). *Yapılandırıcı öğrenme teorisine dayanan etkili bir öğretim yöntemi: Tahmin et-gözle-açıkla-"buz ile su kaynatılabilir mi?[An affective teaching way depend on the theory of constructivist learning: Guess-observe-explain-'can an ice be heated with water'].* A Proceeding presented in V. National Science and Math Education Congress, Ankara.

Kurnaz, M. A., Tarakçı, F., Saydam, A., & Pektaş, M. (2013). Elektriklenme, yıldırım ve şimşek ile ilgili öğrenci zihinsel modellerinin incelenmesi [An analysis of high school students' mental models of electrification, thunder and lightning]. *Uşak University Journal of Social Sciences, 6(*4), 33-51.

Liew, C. W. (2004). *The effectiveness of predict-observe-explain technique in diagnosing students' understanding of science and identifying their level of achievement.* Unpublished Ph.D Thesis, Curtin University of Technology Science and Mathematics Education Centre, Australia.

Mısır, N. (2009). *Elektrostatik ve elektrik akımı ünitelerinde TGA yöntemine dayalı olarak geliştirilen etkinliklerin uygulanması ve etkililiğinin incelenmesi [Application and investigation of the effectiveness of the activities based on the POE method in the units of "electrostatic" and "electric current"].* MA Dissertation. Karadeniz Technical Universty, Trabzon.

Nakhleh, M. B., & Krajcik, J., S. (1994). Influence of levels of information as presented by different technologies on students' understanding of acid, base, and ph concepts. *Journal of Research in Science Teaching, 34*(10), 1077-1096. https://doi.org/10.1002/tea.3660311004

Osborne, R., & Freyberg, P. (1985). *Learning in science: The implication of childrens' science*. Auckland: Heinmann.

Özdemir, A. M. (2012). *İlköğretim 5. sınıf fen ve teknoloji dersi ünitelerinde kavramsal değişim yaklaşımının öğrenci başarısına etkisinin incelenmesi* [*Examining of the effectiveness of conceptual change approach on students? achievement at elementary school fifth-grade science and technology course themes].* PhD Dissertation, Gazi University Institute of Educational Sciences, Ankara.

Özdemir, H. (2011). *Tahmin et-gözle-açıkla stratejisine dayalı laboratuvar uygulamalarının fen bilgisi öğretmen adaylarının asitler-bazlar konusunu anlamalarına etkisi [Effect of laboratory activities designed based on "Predict-Observe-Explain (POE)" strategy on pre-sevice science teachers' understanding of acid-base subject].* MA Dissertation. Pamukkale University, Denizli.

Özmen, H. & Karamustafaoğlu, O. (Ed.) (2019). *Eğitimde araştırma yöntemleri*, Ankara: Pegem Academy.

Palmer, D. (1995). The POE in the primary school: An evaluation. *Research in Science Education, 25*(3), 323-332. https://doi.org/10.1007/BF02357405

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. Newbury Park London New Delhi: Sage Publications.

Senemoğlu, N. (2013). *Gelişim, öğrenme ve öğretim*, (23. Ed.). Ankara: Yargı Publishing House.

TDK. (2019). *Büyük Türkçe sözlük.*

Temizkan, M. (2011). Türkçe öğretmeni adaylarının temel dil becerilerinden okuma ile ilgili kavramları öğrenme düzeyleri ve kavram yanılgıları [The learning levels of teacher candidates about basic concepts of reading skill and misconceptions]. *Dicle Universitesy Journal of Ziya Gökalp Educational Faculty, 17*(2011) 29-47.

Tokcan, H. (2015). *Sosyal bilgilerde kavram öğretimi*. Ankara: Pegem Academy.

Toroslu Çekiç, S. (2011). *Yaşam temelli öğrenme yaklaşımı ile desteklenen 7e öğrenme modelinin öğrencilerin enerji konusundaki başarı, kavram yanılgısı ve bilimsel süreç becerilerine etkisi [Effect of 7e learning model integrated with real-life context based instruction on students' conceptual achievement, misconceptions and science process skills about" energy"].* PhD Dissertation, Gazi University, Ankara.

Töman, U., & Çimer, S. O. (2016). Enerji kavramının farklı öğrenim seviyelerinde öğrenilme durumunun araştırılması [An investigation into the conceptions of energy at different educational levels]. *Journal of Bayburt Education Faculty*, *6*(1), 31-43.

Trochim, W. M. K. (2001). *The research methods knowledge base* (2nd ed). Cincinnati: Atomic Dog Publishing.

Ülgen, G. (2001). *Kavram geliştirme*. (3rd Ed.), Ankara: Pegem Academy.

Yağbasan, R., & Gülçiçek Ç. (2003). Fen öğretiminde kavram yanılgılarının karakteristiklerinin tanımlanması [Description of the characteristics of misconceptions in science education]. *Pamukkale University Journal of Education, 13*, 102-120.

Yel, S. (2015). "Kavram Geliştirme Öğretimi". Öztürk C. (Ed.). *Sosyal bilgiler öğretimi* (p. 111-143). Ankara: Pegem Academy.

Yenilmez, K., & Yaşa, E. (2008). İlköğretim öğrencilerinin geometrideki kavram yanılgıları [Misconceptions of elementary school students in geometry]. *Journal of Uludag University Faculty of Education, 21*(2), 269-290, 463.

Yıldırım, A., & Şimşek, H. (2013). *Sosyal bilimlerde nitel araştırma yöntemleri*. Ankara: Seçkin Publications.

Yılmaz, K., & Çolak, R. (2011). Kavramlara genel bir bakış: Kavramların ve kavram haritalarının pedagojik açıdan incelenmesi [A look at concepts: Investigation of concepts and concept maps from pedagogical perspective]. *Atatürk University Journal of Social Science Institute, 15*(1), 185-204.

Yürümezoğlu, K. Ayaz, S., & Çökelez, A. (2009). İlköğretim ikinci kademe öğrencilerinin enerji ve enerji ile ilgili kavramları algılamaları [Grade 7-9 students' perceptions of energy and related concepts]. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education*, *3*(2), 52-73.