IJATE

Volume 6 Issue 4 2019

International Journal of Assessment Tools in Education

http://ijate.net/

e-ISSN: 2148-7456

IJATE

e-ISSN 2148-7456

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

Volume 6 **Issue 4** **2019**

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

Pamukkale University,
Education Faculty,
Department of Mathematic and Science Education,
20070, Denizli, Turkey

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

Publisher : İzzet KARA
Frequency : 4 issues per year starting from June 2018 (March, June, September, December)
Online ISSN: 2148-7456
Website : http://www.ijate.net/index.php/ijate
http://dergipark.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 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) (Web of Science Core Collection)
- TR Index (ULAKBIM),
- ERIH PLUS,
- DOAJ,
- Index Copernicus International
- SIS (Scientific Index Service) Database,
- SOBIAD,
- JournalTOCs,
- MIAR 2015 (Information Matrix for Analysis of the Journals),
- idealonline,
- CrossRef,
- ResearchBib,
- International Scientific Indexing

Table of Contents

Research Article

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 539–554
https://dx.doi.org/10.21449/ijate.571664

Published at http://www.ijate.net http://submit.ijate.net/en/ *Research Article*

Psychometric Properties of the *School Liking and Avoidance Questionnaire* in a Turkish Preschool Sample

İmray Nur [1,*], Yaşare Aktaş Arnas [2]

[1]Department of Child Development, Osmaniye Korkut Ata University, 80010, Osmaniye, Turkey
[2]Department of Pre-School Education, Cukurova University, 01330, Adana, Turkey

ARTICLE HISTORY

Received: 29 May 2019

Accepted: 08 September 2019

KEYWORDS

School adjustment,
Preschool,
School liking and avoidance questionnaire,
Scale adaptation

Abstract: In this study, the replicability of the factor structure of The School Liking and Avoidance Questionnaire for Turkish preschool children and the psychometric properties of the scale were investigated. The SLAQ consists of 14 items and is used to assess children's emotional and behavioral participation in school based on children's perceptions. 345 children aged 5-6 years were included in the study. The findings of exploratory and confirmatory factor analyses, to assess the validity of the scale in Turkish culture, have identified two distinct but related factors: school liking and school avoidance. This result is consistent with the original scale and the adaptation studies in different cultures. However, there were no significant correlations between children and teacher reports on school liking and school avoidance subdimensions. Cronbach's Alpha and test-retest ratios for subdimensions showed adequate psychometric properties. This data support the Turkish version of the SLAQ as a valid and reliable tool, similar to the original version.

1. INTRODUCTION

Preschool education institutions are considered to be one of the most important contexts in which children can acquire and develop social and early academic skills. Early experiences of children in school constitute the basis for their future social and academic achievements (Fredericks, Blumenfield, Friedel, & Paris, 2005; Ladd & Burgess, 2001; Ladd, Herald, & Kochel, 2006). However, when children start school, they have to cope with many problems (Murray, Murray, & Waas, 2008). Academic challenges, compliance with class and school rules, meeting teacher expectations and being accepted by their peers are some of these problems (Ladd, 1990; Ladd & Price, 1987; Olson & Rosenblum, 1998). In addition, children have to cope with interpersonal problems and cognitive tasks that are becoming increasingly complex throughout the school year. Many children cannot overcome these problems and have many problems related to school adjustment (Rimm-Kaufman, Pianta, & Cox, 2000). Research shows that children coping with these problems and adjustment to school environment are related to positive or negative results. Children develop many ideas, beliefs and attitudes about school in pre-school period. Researchers came to conclusions which show that these ideas,

CONTACT: İmray NUR ✉ imraynur@osmaniye.edu.tr Osmaniye Korkut Ata University, 80010, Osmaniye, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

beliefs and attitudes that children develop about school continue throughout their school life (Fantuzzo, Bulotsky, McDermott, Mosca, & Lutz, 2003; Ladd, Buhs, & Seid, 2000; Ladd & Burgess, 2001; Ladd & Dinella, 2009; Pears, Kim, Capaldi, David, & Fisher, 2012; Pianta & Steinberg, 1992; Smith, 2011). For example, it was found that there is a close correlation between the degree of adjustment to the school perceived by the teachers (positive attitude towards school) and a subsequent (later) higher academic achievement, between positive feelings towards the school in the kindergarten, and high literacy scores in the 5th grade (Hauser-Cram, Durand, & Warfield, 2007), between liking school in pre-school period and reading and mathematics attainments in 8th grade (Ladd & Dinella, 2009). Many studies emphasize that there is a close correlation between school adjustment and school achievement (Birch & Ladd, 1997; Ladd, 1990; Ladd, Kochenderfer, & Coleman, 1996; Ladd & Price 1987; Ladd et al., 2000). Children will probably benefit more from these experiences when they like school and participate in class activities. As opposed to this, when the children have a negative attitude towards the school or avoid the school, this situation will constitute an obstacle for their progress (Ladd, 1990). Ladd and colleagues (2000) stated in their study that children who express that they like school often participate in class activities and show high success. All of these results suggest that adjustment to school in the pre-school period is a subject that should be carefully considered as a developmental problem.

Adjustment to school is a multi-faceted concept with subdimensions such as liking school, avoiding school, academic achievement and dropping-out of school (Goldberg, 2006). In general, school adjustment refers to children's commitment to school or school-related activities, their participation and interest, and their level of being comfortable and successful (Ladd, 1996; Ladd, Buhs, & Troop, 2002). There are many factors that affect children's adjustment to school. Many variables of the child such as child's temperament (Al-Hendawi 2010; Lengua, Wolchik, Sandier, & West, 2000; Morris et al., 2002; Yoleri, 2014), self-esteem (Kaya & Akgün 2016), cognitive readiness and intelligence (Pianta & McCoy, 1997; Reynolds, 1991), executive functions (Sasser, Bierma, & Heinrichs, 2015), self-regulation skills (Morrison, Ponitz, & McClelland, 2010; Williams, Nicholson, Walker, & Berthelsen, 2016) social skills and behavioral problems (Chen, Rubin, & Li, 1995; Ladd, 2006; Ladd & Burgess, 2001) academic skills (Ladd et al., 2000) have been extensively studied by researchers. In addition, factors such as, gender (O'Connor & McCartney, 2007; Yoleri, 2014), socioeconomic level (Ackerman, Brown, & Izard, 2004; Duncan & Brooks-Gunn, 2000), race/ethnicity (Downer, Goble, Myers, & Pianta, 2016), parenting style (Myers, 2007), the quality of the preschool education program (Pianta, La Paro, Payne, Cox, & Bradley, 2002), class climate (Carson & Templin, 2007; Hamre & Pianta, 2001; Jennings & Greenberg, 2009; Robinson, 2013) activities to support adjustment to school employed by teachers and parents (Copeman-Petig, 2015; LoCasale-Crouch, Mashburn, Downer, & Pianta, 2008; Schulting, Malone, & Dodge, 2005) and family participation (Anguiano, 2004; Copeman-Petig, 2015) have been considered by researchers.

Another important factor that affects children's adjustment to school is social relations at school. In this context, relations with teachers and peers have the potential to affect all school experiences of children. Specifically, children's relationships with their teachers provide a significant contribution to, their relationships with their peers, social competencies, school adjustment and to the development of their academic skills (Baker, Grant, & Morlock, 2008; Birch & Ladd, 1997; Doumen, Koomen, Buyse, Wouters, & Verschueren, 2012; Gallagher, 2015; Howes, Phillipsen, & Peisner-Feinberg, 2000; Pianta & Stulhman, 2004; O'Connor & McCartney, 2007). Similarly, while contributing to their learning of social and cognitive skills (Hartup & Stevens, 1997; Howes, 1996) the quality of children's relationships with their peers is closely related to their ability to adapt to the school (Ladd, 1990; Maguire & Dunn, 1997). Because quality friendship gives children more affirmation, support, sincerity and confidence.

For example, Howes, Rubin, Ross, and French (1988) stated that children who have friends have a higher adjustment to school transitions than those with no friends. Similarly, Ladd (1990) emphasized that children who started preschool education had better adjustment to school if they had classmates from their earlier years (from nursery), and if these friendships are continued throughout the year these children gain more positive impressions about the school.

Although the current literature provides a substantial foundation on adjustment to school and the factors affecting adjustment to the school, in many instances, teachers' perceptions have been considered as indicators of adjustment to school (Birch & Ladd, 1997; Buhs, Ladd, & Herald, 2006; Claes, 2010; Engle, McElwain & Lasky, 2011; Goldberg, 2006; Myers, 2007; Sette, Hipson, Zava, Baumgartner, & Coplan, 2018). Teachers' perspectives provide important information when it comes to children's adjustment to school, but they cannot replace information obtained from children (Smith, 2011). At the same time, children are valid and very valuable sources of information about their internal processes and problems. Children are experts in their own lives, they are aware of their experiences as learners and have the ability to express their ideas (Daly et al., 2007; Einarsdottir, 2005; Kragh-Muller & Isbell, 2011). In addition, having information about children's feelings and thoughts and taking their point of view into consideration are important for preventing emergence of problems related to school adjustment and all subsequent risk factors. Given this situation, the researchers worked on determining the preschool children's thoughts about the school based on their perceptions. The most important of these efforts was The School Liking and Avoidance Questionnaire ([SLAQ], Ladd et al., 2000), which was adapted from the studies by Ladd and Price (1987) and Ladd (1990) to determine the attitudes of children towards school based on children's perceptions.

The SLAQ, designed to assess children's feelings about school, consists of 14 items. The items require children to express their positive feelings towards school (9 items, school liking), and feelings such as the desire to go home from school (5 items, school avoidance). There are many studies examining psychometric properties of SLAQ in different cultures and different age groups. In a study conducted by Zhang (2016) with Chinese children (M_{age} = 14.25), exploratory factor analysis did not support the factor structure of the original scale. The school liking subscale included 4 items (Cronbach's α = .76 urban group, .61 rural group) and the school avoidance subscale included 4 items (Cronbach's α = .82 urban group, .77 ruralgroup). SLAQ's Italian validity and reliability study was carried out with children with the average age of 7 years, 7 months. In the Italian SLAQ, school liking subscale included 8 items (Cronbach's α = .89) and the school avoidance subscale included 5 items (Cronbach's α = .82) (Tomada, Schneider, de Domini, Greenman, & Fonzi, 2005). In another study with Greek first-grade students (M_{age} = 77.22 months), Cronbach's alphas for school liking were .84, .86, and .89, for school avoidance were .75, .76 and .79 at pre-, post- , and follow-up assessments respectively (Vassilopoulos, Brouzos, & Koutsianou, 2018). SLAQ was also adapted for Japanese children by Otsui and colleagues but the results could not be reached because the article written in Japanese (cited in Honma & Uchiyama, 2014). In the validity and reliability study conducted by Smith (2011) with American children aged 5-12 years, the school liking subscale contained 7 items (Cronbach's α = .89 -.91) and the school avoidance subscale contained 5 items (Cronbach's α = .79 -.84).

In Turkey, although some studies were conducted on preschool children's adjustment to school their number is quite limited. In these studies, factors which are thought to be effective in children's adjustment to school such as, gender and problem behaviors (Yoleri, 2015), peer relations and peer acceptance (Gülay & Erten, 2011; Yoleri, 2015), social skills (Gülay, 2011), mother attitudes (Gülay-Ogelman, Önder, Seçer, & Erten, 2013), relationships with mother and teacher (Nur, Aktaş-Arnas, Abbak, & Kale, 2018) were investigated. However, these studies were based only on teacher perceptions. It is usual to ask teachers about children's feelings towards school because teachers spend long hours with children and they collect a lot of information about children during the day. However, in the literature, there are some concerns about collecting data on children from adults only. One concern is that teachers have the potential to act biased as raters (Kesner, 2000; Saft & Pianta, 2001). In addition, discrepancies found in some studies based on child, parent and teacher reports, between the perceptions of children and adults have increased these concerns. For example, some studies have emphasized that teachers' and children's perceptions on adjustment to school and teacher-child relationship have little or no coherence (Harrison, 2004; Murray et al., 2008; Smith, 2011). In another study, it was found that peer reports gave more accurate results than parents and teacher reports (Clements, Musci, Leoutsako, & Ialongo, 2015). For this reason, it is an important matter to collect child reports in studies related to children.

Children develop and learn in environments where they receive attention and are valued and happy. In environments where their thoughts and feelings are listened and accepted, children feel safe and find opportunities to learn and explore. Knowing the feelings of children about school from their point of view will make it possible to respond to their needs both individually and as a group. This study was planned to determine whether the factor structure of SLAQ is replicable for 5-6 years old Turkish children and to evaluate the psychometric properties of the scale. If it is confirmed that the scale is valid and reliable in Turkish culture, the need for a valid tool that researchers can employ aiming to evaluate the emotions of children about the school will be satisfied. Thus, it is thought that adaptation of SLAQ to Turkish will make a significant contribution to the field.

2. METHOD

2.1. Participants

The study groups were formed by convenience sampling method from 5-6 years old children attending pre-school education institutions in central Osmaniye in southern Turkey. Firstly, schools were visited and the purpose of the research was explained to the administrators and teachers. Afterwards, the parents of the children from the classrooms of the 26 teachers who agreed to participate in the research were sent an informed consent form. Children whose parents gave their informed consent were included in the study.

Since this research was an adaptation study, two different study groups were formed. A total of 345 children (M_{age} = 67.10 months, *SD* = 4.10, range 54 to 77), 161 females and 184 males, participated in the study. To examine the factor structure of the scale, exploratory factor analysis was performed with data collected from 129 children (63 female, 66 male, M_{age} = 66.59 months, *SD* = 3.30, *range* 60 to 73). In addition, confirmatory factor analysis was performed with data collected from 216 children (98 female, 118 male, M_{age} = 67.41 months, *SD* = 4.50, *range* 54 to 77).

2.2. Instruments

2.1.1. *School liking and avoidance questionnaire*

School Liking and Avoidance Questionnaire (SLAQ) was adapted from studies by Ladd and Price (1987) and Ladd (1990) to determine children's attitudes towards school. Constituting of a total of 14 items, the SLAQ has a self report measure and a two-factor structure. The school liking subdimension (Items:1, 2, 4, 6, 7, 8, 10, 11, 12) assesses children's positive perceptions and feelings about the school (Three reverse score items, Is school a fun place to be? Do you like being in school?) and school avoidance subdimension (Items:3, 5, 9, 13, 14) assesses children's school avoidance desire (Do you wish you didn't have to go to school? Do you wish you could stay home from school?). During individual interviews with children, children are asked to evaluate the items with a three-point scoring (“yes,” “sometimes,” or “no,” which were scored as 3, 2, and 1 respectively). For each subdimension, the total score is calculated taking the mean scores for each item. High scores for school liking subdimension indicate positive feelings about the school, while high scores for the school avoidance subscale indicate a higher desire for school avoidance. Ladd and colleagues (1996) calculated the internal consistency coefficients for subdimensions for the fall and spring periods separately in their study and reported that the coefficients were strong (School liking Cronbach's α fall= .87, spring = .91, school avoidance Cronbach's α fall=.76, spring = .81).

2.1.2. *Teacher rating scale of school adjustment*

Teacher Rating Scale of School Adjustment (TRSSA) has been developed by Ladd, Kochenfender and Coleman (1996) to evaluate children's school adjustment skills based on teachers' perceptions. The scale consists of 4 subdimensions (school liking, cooperative participation, school avoidance and self-directedness). The 5-item school liking subscale determines the teacher's perception of how much the child likes school (Cronbach's α = .89). The cooperative participation subscale, which consists of 8 items, measures the extent to which the child accepts the teacher's authority, classroom rules and responsibilities (Cronbach's α = .92). The school avoidance subscale consists of 5 items and based on teacher perceptions aims to determine the degree to which the child avoids the classroom environment (Cronbach's α = .74). The 9-item self-directedness subscale evaluates the child's independent or self-directed behavior within the classroom (Cronbach's α = .91). Each item in the scale is evaluated using a 3-point likert scale (from 0 = doesn't apply to 3 = certainly applies) (Birch & Ladd, 1997). The internal consistency coefficient of the scale, which was adapted to Turkish by Önder and Gülay (2010), was found to be .70 for the whole scale. The internal consistency coefficients of the subscales ranged between .67 and .84 (Önder & Gülay, 2010).

Within the scope of the current study, in order to test the internal consistency of TRSSA, Cronbach's alpha values of the subscales were examined. As a result of the analysis, the alpha value was determined as .87 for school liking subscale, .86 for cooperative participation, .87 for school avoidance subscale, and .76 for self-directedness subscale.

2.1.3. *Preschool and kindergarden behavior scales*

Preschool and Kindergarden Behavior Scales (PKBS-2) was developed by Kenneth W.Merrell in 1994 to evaluate social skills and problem behaviors of 3-6 year old children. In 2003, the scale was revised and a norm study was conducted with 3,317 children aged between 3 and 6 years (Merrell, 2003). The scale consists of two independent scales: Social Skills and Problem Behavior scales. Scale is a 4-point Likert type scale. The validity and reliability study of the scale for Turkish children was done by Özbey (2009). Social Skills Scale consists of three subdimensions; Social Cooperation (11 items), Social independence and Social acceptance (8 items) and Social Interaction (4 items) with a total of 23 items. The Cronbach's Alpha values of the Social Skills Scale subscales were .92, .88, .88, and the overall Cronbach's Alpha Scale of the Social Skills Scale was .94, respectively (Özbey, 2009). The high total score indicates that children have high social skills. The Problem Behavior Scale consists of four factors: Externalizing Problems (16 items), Internalizing Problems (5 items), antisocial (3 items) and egocentric (3 items). The Cronbach's Alpha values of the Problem Behavior Scale subdimensions were .95, .87, .81, .72, and the overall Cronbach's Alpha value of the Problem Behavior Scale was .96 (Özbey, 2009).

In order to test the internal consistency of the scales in the present study, Cronbach's Alpha values were examined. The Cronbach's Alpha values of the Social Skills Scale subscales were .92, .86, .96, and the Cronbach's Alpha value of the Social Skills Scale was .92, respectively. The Cronbach's Alpha values of the Problem Behavior Scale subdimensions were .96, .81, .65, .74, and the overall Cronbach's Alpha value of the Problem Behavior Scale was .93.

2.3. Procedure

The adaptation of the School Liking and Avoidance Scale was done through translation-re-translation study. For this purpose, the 14-item scale was translated into Turkish by researchers and two English experts. The suitability of the translation by the researchers was tested by comparing with the other two translation studies. Translated scale was sent to three academics who are experts in preschool education and who have mastery of English, for correction and assessment of how closely the items represent the original content. Corrections have been made by taking into consideration the suggestions made for using the synonyms of some words. The corrected translation of the scale was sent to four academicians specialized in pre-school education in order for assessment of the comprehensibility and fitness for purpose. The Turkish version of the scale, which was revised in line with the suggestions, was translated back to English. Afterwards, translations were reviewed for semantic shifts. Thus, it was concluded that the Turkish version of the scale was ready for implementation for the validity and reliability studies. Following these procedures, a pilot run was conducted with 19 children to evaluate the scale's comprehensibility.

Given the developmental characteristics of pre-school age children and the fact that they don't read or write, researchers have suggested that different scale designs are required to minimize situations that may interfere with young children's responses (Harter & Pike, 1984; Lewis & Lindsay, 2000; Zhang, Smith, Lam, Brimer, & Rodriquez, 2002). One of these is the addition of pictorial representation to scale items in order to facilitate the response of children (Hanna, Risden, Czerwinski, & Alexander, 1999; Mantzicopoulos, French, & Maller, 2004). Harter and Pike (1984) suggest that the visual image matching in the scales encourages a meaningful understanding of the content and facilitates the production of meaningful responses as it ensures children's attention and participation. One of the most commonly used pictorial representations when working with young children is the "smileyface" with facial expressions ranging from unhappy to happy (Reynolds-Keefer, Johnson, Dickenson, & McFadden, 2009; Hall, Hume, & Tazzyman, 2016). In this study, for the Turkish sample, "sad" facial expression for "no" answer, "neutral" facial expression for "sometimes" answer, and "happy" facial expression for "yes" answer were added next to the items in the scale (☹.😐.☺). Children were encouraged to paint or mark these facial expressions according to their answers.

The data of the study were collected by one of the researchers through individual interviews with children. She had spent at least three hours in the children's classrooms in order to be acquainted and build rapport with the children before interviews started. During this period, the researcher introduced herself, participated in the activities of children and played with them. Later, the children whose parents gave their consent were individually invited to the interview. The researcher and the child met in a private room at children's school, and the researcher told the child about the research and asked for her/his consent to participate also. Children were informed that their answers will remain confidential and they can stop responding at any time. All children have agreed to participate.

The final Turkish version of the scale was applied to 345 preschool children in order to evaluate the validity and reliability of the scale in Turkish sample. Exploratory and confirmatory factor analyses were performed on the data transferred to the computer. In the confirmatory factor analysis, x^2/sd, Root Mean Square Error of Approximation (RMSEA), Non-Normed Fit Index (NNFI), Goodness of Fit Index (GFI), Comparative Fit Index (CFI) and Incremental Fit Index (IFI) indices were used to assess the validity of the model fit. In order to determine Concurrent Validity, Pearson correlation coefficients between TRSSA and PKBS–2 completed by teachers for 107 children and SLAQ were calculated. The reliability of the scale was examined with Cronbach's Alpha internal consistency coefficient. In addition, in order to evaluate how consistently the instrument measures, second interviews were performed with 97 children in three weeks intervals and the correlation between the data obtained was examined.

3. RESULT / FINDINGS

3.1. Construct Validity

3.1.1. Exploratory factor analysis

To examine the factor structure of the School Liking and Avoidance Questionnaire, data gathered from 129 children were included in the Exploratory Factor Analysis (EFA). In determining the items to be included in the scale in exploratory factor analysis, it was stipulated that the eigenvalues of the items should be at least 1.00, the factor load values of the items should behigher than .40, the items should be contained in a single factor, and there should be a minimum difference of 0.10 between the factor loads of the items loaded in two different factors. The suitability of the data collected from School Liking and Avoidance Scale for factor analysis was evaluated by the Kaiser-Meyer-Olkin (KMO) coefficient and Barlett's test. Kaiser-Meyer-Olkin's sample adequacy measure (KMO = 0.85) and Bartlett's test of sphericity ($\chi2$=1505.535; p<.0001) results showed that the data were suitable for factor analysis.

Table 1. *Results of the Principal Component Analysis of the School Liking and Avoidance Questionnaire*

	Item	1st Factor	2nd Factor	Corrected Item-Factor Correlation
SLAQ	Item 11	.867		.78
	Item 7	.848		.82
	Item 8	.836		.77
	Item 10	.766		.71
	Item 6	.741		.73
	Item 12	.712		.67
	Item 4	.684		.64
	Item 1	.677		.80
	Item 2	.511		.39
	Item 14		.884	.77
	Item 3		.817	.78
	Item 5		.774	.77
	Item 9		.639	.60
	Exp. Variance %	40.877	27.312	

Factor analysis was performed twice on the data obtained from SLAQ. As a result of principal components factor analysis using the Varimax rotation method for the 1st factor analysis, 2 factors with eigenvalues greater than 1 were obtained similar to the original scale. These 2 factors explain 63.62% of the total variance. For Item 13 of the scale (Do you feel more happy when you go home from school?) item load value is determined as 0.19. Since the item load value was below 0.40, this item was removed from the scale and the remaining 13 items were re-analyzed. The results obtained from exploratory factor analysis are given in Table 1.

With the new factor analysis, a two-factor and 13-item structure explaining 68.190% of the total variance was obtained. The first factor is school liking subdimension consisting of items 1, 2, 4, 6, 7, 8, 10, 11 and 12 (the items 2, 6 and 12 are calculated in reverse). The second factor is school avoidance subdimension consisting of items 3, 5, 9 and 14. It was determined that the item load values of the scale were between .51 and .88. When the lower limit for the factor load value is taken as .32 as stated in the scale development and adaptation studies, it can be said that the factor load values of the scale are sufficient (Büyüköztürk, 2002; Kline, 2005). In order to examine the validity of the items, the total correlation values of the items were found to be between .39 and .82. Item total correlation values of .30 and above is considered to be sufficient to distinguish the property to be measured. According to the data obtained, each item included in the scale was highly correlated with the total score of the scale and met item validity.

3.1.2. Confirmatory factor analysis

Confirmatory factor analysis (CFA) was performed twice to examine the model fit of the 14-item two-factor structure of SLAQ. The analyzes were carried out on the data obtained from the second sample ($N = 216$). When the confirmatory factor analysis was performed, at first the criteria for the model fit were examined. In the first CFA analysis the following values were found, $\chi 2/df$ =2.548 (p <.001), root mean square error approximation (RMSEA) .08, non-normed fit index (NNFI) .92, goodness of fit index (GFI) .90, comparative fit index (CFI) .93 and incremental fit index (IFI) .93. However, it was found that the 13th item was not significantly predicted by the second factor. The analysis was repeated after eliminating this item. In the CFA, conducted in order to examine the model fit of the 13-item two-factor structure of SLAQ, the chi-square test was significant ($\chi 2$ = 143.397, df= 62, χ2/df =2.313 p <.001) and indicated that the model was a good fit. In addition, as shown in Table 2, the fit indexes determined by using maximum likelihood estimation indicate that the model is a good fit. RMSEA = .07, NNFI = .93, GFI = .91, CFI = .94 and IFI = .94. These results obtained by eliminating one item from the scale (Item 13) indicate that the factor structure of the model has been verified and that the original structure of SLAQ is suited for use for Turkish culture.

Table 2. *Summary of Fit Indices from Confirmatory Factor Analysis*

$\chi 2$	df	$\chi 2$/df	RMSEA	NNFI	GFI	CFI	IFI
143.397	62	2.313	0.7	0.93	0.91	0.94	0.94

In order to evaluate the construct validity of School Liking and Avoidance Scale, correlations between factors were also taken into consideration. Correlation (r = −625, p < .01) between school liking subdimension (X = 25.5, ss = 2.9) and school avoidance subdimension (X = 4.77, ss = 1.6) showed a significant negative correlation between the subdimensions of the scale and there is no multiple correlation problem.

3.1.3. Concurrent validity

In order to evaluate the Concurrent Validity, the correlations between School Liking and Avoidance Scale, Teacher Rating Scale of School Adjustment, Preschool and Kindergarden Behavior Scales (Social Skills and Problem Behavior Scales) were examined. There are no significant correlations between the perceptions of children and the perceptions of teachers about school liking and school avoidance. Similarly, the correlations between the school liking and avoidance based on children's perceptions and each subdimension of Social Skills Scale based on teacher perceptions are not significant. However, significant negative correlations were identified between the school liking subdimension based on children's perceptions and antisocial (r = -.322, p <.001) and egocentric (r = -.268, p <.001) subdimensions of the Problem Behavior Scale based on teachers' perceptions.

3.2. Reliability Analysis

SLAQ reliability was measured by Cronbach's Alpha coefficient and test-retest methods. The Cronbach's Alpha Coefficient for the SLAQ school liking subdimension is .92 and .87 for the school avoidance subdimension. For test-retest reliability the correlation between measurements made in three-week intervals was examined. High levels of significant correlations between two applications were identified for school liking (r = .86; p <0.01) and avoidance (r = .84; p <0.01) subdimensions. This result shows that the evaluation tool gives stable results over time.

4. DISCUSSION

In this study, the replicability of the factor structure of The School Liking and Avoidance Questionnaire for Turkish preschool children and the psychometric properties of the scale were examined.SLAQ is used to assess children's emotional and behavioral involvement in school. The EFA and CFA findings, which were conducted to evaluate the construct validity of the scale in Turkish culture, identified two distinct but related factors: school liking and school avoidance. This result is consistent with the original scale and the adaptation studies in different cultures (Honma & Uchiyama, 2014; Ladd et al., 2000; Smith, 2011; Tomada et al., 2005; Vassilopoulos et al., 2018).

The relationship between child and teacher reports was examined to evaluate the construct validity of the School Liking and Avoidance Questionnaire subscales. There were no significant correlations between the perceptions of children and the perceptions of teachers on school liking and school avoidance. Similar results were obtained in previous studies (Murray et al., 2008; Smith, 2011). Murray and colleagues (2008) reported in their study on teacher-child relationship and children's school adjustment that there was a low correlation between the children's and the teachers' perceptions and additionally the children's perceptions of teacher-child relationship were more predictive of school liking. In another study by Harrison (2004), the correlation between children's reports of school liking and avoidance and teachers' reports of task orientation and assertive social skills differ with regard to girls and boys. There was no significant correlation between the scores of girls' school liking and avoidance scores and teacher scores. In fact, a high correlation is expected between teacher reports and child reports. Because teachers are together with children all day and have the opportunity to observe them closely. Although child and teacher reports of school liking and avoidance subdimensions are evaluating the same structure, some limitations may have led to the development of this situation. For example, in crowded classrooms, teachers may have overlooked the behavior of other children as they paid attention to children who were over-exhibiting school avoidance behavior (Smith, 2011). The expectations of teachers and children from the school may be different. Teachers may perceive children who participate in activities without any problems and have academic success as liking the school and they may not be observing adequately. In fact, according to the social discipline model of Dreikurs, at the root of children's behaviour lies the desire to belong to a group, a class. According to Drekurs, what motivates the children to misbehave can be the subconscious desire to draw attention, seek power or take revenge (Sadık, 2018). In this sense, children who want to attract the attention of teachers or peers by displaying such negative behaviors may not be desiring to get away from classroom or avoid school, instead they may be desiring to belong and be acknowledged. This desire can be the basic impulse underlying their negative behavior. However, teachers may think that these children avoid school.

Many studies based on teachers' perceptions emphasize the relationship between social skills and behavioral problems of children and school adjustment (Chen et al., 1995; Ladd, 2006; Ladd & Burgess, 2001). However, different findings have been reached in studies focusing on the correlations between children's perspective on school and the teachers' perpective on social skills and problem behaviour. In his study, Huang (2010) stated that there are correlations between school liking and avoidance based on children's perceptions and positive social behaviors and problem behaviors based on teachers' perceptions. In this study, no significant correlation between the school liking and avoidance scores based on the children's perceptions and the social skills scores based on teachers' perceptions was identified. However, the correlations between the liking school subdimension based on children's perceptions and antisocial and egocentric behavior problems based on teachers' perceptions are significant. There are studies that support this finding in the literature. For example, Lee (2014), in his study, stated that there were correlations between externalizing problems based on teachers' perceptions and liking school, however there were no correlations between internalizing problems and children's reports. Similarly, Harrison (2004) concluded that there are correlations between boys' school liking and avoidance children's reports and acting out behaviors teachers' reports, but there is no correlation with being shy/anxious for both boys and girls.

Looking at the results of the internal consistency coefficients examined in order to evaluate the reliability of the scale, and considering the accepted reliability level of the measurement tools that can be used in the research is .70, the scale can be declared reliable. This result supports the results of the study by Smith (2011) (School liking Cronbach's α = .91, school avoidance Cronbach's α = .80). The results of the test and re-test to evaluate the reliability show that the scale gives stable results over time. In summary, the validity and reliability results for SLAQ show that the scale is valid and reliable in Turkish culture.

5. CONCLUSION

The findings of this study show that the Turkish version of The School Liking and Avoidance Questionnaire gives valid and reliable scores for Turkish preschool children. It is thought that the tool will contribute to the researchers who want to assess school adjustment problems, to identify children who do not like school or avoid school, to take measures for children at risk, and to investigate the correlation between the degree of emotional and behavioral participation in school and other adjustment indicators. There is a need for measurement tools to evaluate the school adjustment of Turkish preschool children based on children's perceptions and this scale adaptation is expected to fill this gap in the field. In addition, the study of the validity and reliability of the tool in different cultures will provide the basis for intercultural studies. In this context, the study may also contribute to the international literature.

Future research should examine the validity and reliability of SLAQ with participants at varying age groups and socioeconomic levels. Research conducted in different regions of Turkey, in varying types of schools and with larger samples may provide further evidence for the validity and reliability of SLAQ. In this study, because there is no other measurement tool based on the perceptions of children to evaluate the school adjustment of children, concurrent validity was evaluated based on the perceptions of teachers. It may be beneficial to employ more social based criteria, such as peer relations and teacher-child relations based on children's perceptions, presumed to affect school adjustment.

ORCID

İmray Nur http://orcid.org/0000-0002-1905-1655

6. REFERENCES

Ackerman, B. P., Brown, E. D., & Izard, C. E. (2004). The relations between contextual risk, earned income, and the school adjustment of children from economically disadvantaged families. *Developmental Psychology,40,* 204-216. doi: 10.1037/0012-1649.40.2.204

Al-Hendawi, M. (2010). *The predictive relationship between temperament, school adjustment, and academic achievement: A 2-year longitudinal study of children at-risk.* (Doctoral dissertation**)**. Retrieved from https://scholarscompass.vcu.edu/etd/2276/

Anguiano, R. P. V. (2004). Families and schools: The effect of parental involvement on high school completion. *Journal of Family Issues, 25*(1), 61 - 85. doi: 10.1177/0192513X03256805

Baker, J. A., Grant, S., & Morlock, L. (2008). The teacher-student relationship as a developmental context for children with internalizing or externalizing behavior problems. *School Psychology Quarterly, 23*(1), 3-15. doi: 10.1037/1045-3830.23.1.3

Birch, S. H., & Ladd, G. W. (1997). The teacher-child relationship and children's early school adjustment. *Journal of School Psychology, 35,* 61-79. doi: 10.1016/S0022-4405(96)00029-5

Buhs, E. S., Ladd, G. W., & Herald, S. L. (2006). Peer exclusion victimization: Processes that mediate the relation between peer group rejection and children's classroom engagement and achievement. *Journal of Educational Psychology, 98,* 1-13. doi: 10.1037/0022-0663.98.1.1

Büyüköztürk, Ş. (2002). Factor analysis: Basic concepts and using to development scale. *Educational Administration in Theory & Practice, 32,* 470-483.

Carson, R. L., & Templin, T. J. (2007). *Emotion regulation and teacher burnout: Who says that the management of emotional expression doesn't matter?* Paper presented at the American Education Research Association Annual Convention, Chicago.

Chen, X., Rubin, K. H., & Li, B. (1995). Depressed mood in Chinese children: Relations with school performance and family environment. *Journal of Consulting and Clinical Psychology, 63,* 938–947. doi: 10.1037/0022-006X.63.6.938

Claes, B. (2010). *Transition to kindergarten: The impact of preschool on kindergarten adjustment.* **(**Doctoral dissertation**).** Retrieved from https://aura.alfred.edu/bitstream/handle/10829/2666/Claes%2C%20Bethany%202010.pdf?sequence=1&isAllowed=y

Clemans, K. H., Musci, R. J., Leoutsakos, J. M. S., & Ialongo, N. S. (2014). Teacher, parent, and peer reports of early aggression as screening measures for long-term maladaptive outcomes: Who provides the most useful information? *Journal of Consulting and Clinical Psychology, 82*(2), 236-247. doi: 10.1037/a0035918

Copeman-Petig, A. M. (2015). *The transition to kindergarten: factors associated with a positive adjustment.* **(**Doctoral dissertation**)**. Retrieved from https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=5804&context=etd

Daly, M., Forster, A., Murphy, R., Sweeney, A., Brennan P., & Maxwell, M. (2007). Children's voices in the Framework for Early Learning –A portraiture study. *The OMEP Ireland Journal of Early Childhood Studies, 1,* 57-71.

Doumen, S., Koomen, H. M. Y., Buyse, E., Wouters, S., & Verschueren, K. (2012). Teacher and observer views on student–teacher relationships: Convergence across kindergarten and relations with student engagement. *Journal of School Psychology, 50*, 61-76. doi: 10.1016/j.jsp.2011.08.004

Downer, J. T., Goble, P., Myers, S. S., & Pianta, R. C. (2016). Teacher-child racial/ethnic match within pre-kindergartenclassrooms and children's early school adjustment. *Early Childhood Research Quarterly, 37,* 26–38. doi: 10.1016/j.ecresq.2016.02.007

Duncan, G. J., & Brooks-Gunn, J. (2000). Family poverty, welfare reform, and child development. *Child Development, 71,* 188–196. doi: 10.1111/1467-8624.00133

Einarsdottir, J. (2005). We can decide what to play! Children's perception of quality in an Icelandic play school. *Early Education and Development*, *16*(4), 469-488. doi: 10.1207/s15566935eed1604_7

Engle, J. M., Mcelwain, N. L., & Lasky, N. (2011). Presence and quality of kindergarten children's friendships: Concurrent and longitudinal associations with child adjustment in the early school years. *Infant and Child Development*, *20*(4), 365-386. doi: 10.1002/icd.706

Fantuzzo, J. W., Bulotsky, R., McDermott, P., Mosca, S., & Lutz, M. N. (2003). A multivariate analysis of emotional and behavioral adjustment and preschool educational outcomes. *School Psychology Review, 32*(2), 185-203.

Fredericks, J. A., Blumenfeld, P., Friedel, J., & Paris, A. (2005). School engagement. In K. A. Moore & L. Lippman (Eds.), *Conceptualizing and measuring indicators of positive development: What do children need to flourish* (pp. 305-321). New York: Kluwer Academic/Plenum Press.

Gallagher, E. (2015). *Teacher-student conflict and student aggression in kindergarten.* Retrieved from http://steinhardt.nyu.edu/opus/issues /2014/spring/gallagher

Goldberg, C. A. (2006). *Transitioning to preschool: The role of withdrawn behavioral subtypes and the teacher-child relationship in early school adjustment***.** **(**Doctoral dissertation**).** Retrieved from https://search.proquest.com/openview/da885ed4effc3bdcef591db6c937fa5f/1?pq-origsite=gscholar&cbl=18750&diss=y

Gülay, H. (2011). 5-6 yaş grubu çocuklarda okula uyum ve akran ilişkileri [School adjustment and peer relatıonships of 5-6 years old children]. *Electronic Journal of Social Sciences, 10*(36), 1-10.

Gülay, H., & Erten, H. (2011). Okul öncesi dönem çocuklarının akran kabullerinin okula uyum değişkenleri üzerindeki yordayıcı etkisi [Predicting effect of play behaviors of preschool children on peer relationships]. *E-International Journal of Educational Research, 2*(1), 81-92.

Gülay-Ögelman, H., Önder, A., Seçer, Z., & Erten, H. (2013). Anne tutumlarının 5-6 yaş çocuklarının sosyal becerilerini ve okula uyumlarını yordayıcı etkisi [The predictor effect of attitudes of mothers on the social skills and school adaptation of their 5–6-year-old children]. *Selcuk University Social Sciences Institute Journal*, *29*, 143–152.

Hall, L., Hume, C., & Tazzyman, S. (2016). Five degrees of happiness: Effective smileyface Likert Scales for evaluating with children. In Proceedings of theThe 15th International Conference on Interaction Design and Children - IDC '16, 311–321. https://doi.org/10.1145/2930674.2930719

Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development, 72*(2), 625-638.

Hanna, E., Risden, K., Czerwinski, M., & Alexander, K., J. (1999). The role of usability research in designing children's computer products. In A. Druin (Ed.), *The design of children's technology* (pp. 4-26). San Francisco: Morgan Kaufmann.

Harrison, L. J. (2004). *Do children's perceptions of themselves, their teachers, and school accord with their teachers' ratings of their adjustment to school?* Paper presented to the Australian Association for Research in Educationnational Conference, Melbourne, Nov 29 –Dec 2.

Harter, S., & Pike, R. (1984). The pictorial scale of perceived competence and social acceptance. *Child Development, 55,* 1969-1982. doi: 10.2307/1129772

Hartup, W. W., & Stevens, N. (1997). Friendships and adaptation in the life course. *Psychological Bulletin*, *12*, 355-370. doi:10.1037/0033-2909.121.3.355

Hauser-Cram, P., Durand, T. M., & Warfield, M. E. (2007). Early feelings about school and lateracademic outcomes of children with special needs living in poverty. *Early Childhood Research Quarterly*, *22*(2), 161-172. doi: 10.1016/j.ecresq.2007.02.001

Honma, Y., & Uchiyama, I. (2014). Emotional engagement and school adjustment in late childhood: The relationship between school liking and school belonging in Japan. *Psychological Reports 114*(2), 496-508. doi: 10.2466/21.10.PR0.114k19w7

Howes, C. (1996). The earliest friendships. In W. M. Bukowski, A. F. Newcomb & W. W. Hartup (Eds.), *The Company they keep: Friendships in childhood and adolescence* (pp. 66-86). New York: Cambridge University Press.

Howes, C., Rubin, K. H., Ross, H. S., & French, D. C. (1988). Peer interaction of young children. *Monographs of the Society for Researchin Child Development, 53*(1), 1-92. doi:10.2307/1166062

Howes, C., Phillipsen, L. C., & Peisner-Feinberg, E. (2000). The consistency of perceived teacher–child relationshipsbetween preschool and kindergarten. *Journal of School Psychology, 38*, 113-132. doi: 10.1016/S0022-4405(99)00044-8

Jennings, P. A., & Greenberg, M. T. (2009). The prosocial classroom: Teacher social and emotional competence in relation to student and classroom outcomes. *Review of Educational Research, 79,* 491-525. doi: 10.3102/0034654308325693

Kaya, Ö. S., & Akgün, E. (2016). The study of school adjustment of preschool children in the point of some variables. *Elementary Education Online, 15*(4), 1311-1324. doi: 10.17051/io.2016.51992

Kesner, J. E. (2000). Teacher characteristics and the quality of child–teacher relationships. *Journal of School Psychology, 28,* 133-149. doi: 10.1016/S0022-4405(99)00043-6

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

Kragh-Muller, G., & Isbell, R. (2011). Children's perspectives on their everyday lives in childcare in two cultures: Denmark and the United States. *Early Childhood Education Journal, 39*, 17–27. doi: 10.1007/s10643-010-0434-9

Ladd, G. W. (1990). Having friends, keeping friends, making friends, and being liked by peers in the classroom: Predictors of children's early school adjustment. *Child Development, 61,* 1081-1100. doi: 10.2307/1130877

Ladd, G. W. (1996). Shifting ecologies during the 5–7 yearperiod: Predicting children's adjustment during thetransition to grade school. In A. Sameroff & M. Haith (Eds.), *The five to seven-year shift* (pp. 363–386). Chicago: University of Chicago Press.

Ladd, G. W. (2006). Peer rejection, aggressive or withdrawn behavior, and psychological maladjustment from ages 5 to 12: An examination of four predictive models. *Child Development, 77*(4), 822-846. doi: 10.1111/j.1467-8624.2006.00905.x

Ladd, G. W., Buhs, E. S., & Seid, M. (2000). Children's initial sentiments about kindergarten: Isschool liking an antecedent of early classroom participation and achievement? *Merrill-Palmer Quarterly, 46*(2), 255-279.

Ladd, G. W., Buhs, E. S., & Troop, W. (2002). Children's interpersonal skills and relationships in school settings: Adaptive significance and implications for school-based prevention and programs. In P. K. Smith & C. Hart (Eds.), *Blackwell handbook of childhood social development* (pp. 394-415). Oxford, UK: Blackwell.

Ladd, G. W., & Burgess, K. B. (2001). Do relational risks and protective factors moderate the linkages between childhood aggression and early psychological and school adjustment? *Child Development, 72,* 1579-1601. doi: 10.1111/1467-8624.00366

Ladd, G. W., & Dinella, L. M. (2009). Continuity and change in early school engagement: Predictive of children's achievement trajectories from first to eighth grade? *Journal of Educational Psychology*, *101*(1), 190-206. doi: 10.1037/a0013153

Ladd, G. W., Herald, S. L., & Kochel, K. P. (2006). School readiness: Are there social prerequisites? *Early Education and Development*, *17*(1), 115 - 150. doi: 10.1207/s15566 935eed1701_6

Ladd, G. W., Kochenderfer, B. J., & Coleman, C. C. (1996). Friendship quality as a predictor of young children's early school adjustment. *Child Development, 67,* 1103-1118. doi: 10.2307/1131882

Ladd, G. W., & Price, J. M. (1987). Predicting children's social and school adjustment following the transition from preschool to kindergarten. *Child Development, 58,* 1168-1189. doi: 10.2307/1130613

Lengua, L. J., Wolchik, S. A., Sandier, I. N., & West, S. G. (2000). The additive and interactive effects of parenting and temperament in predicting problems of children of divorce. *Journal of Clinical Child Psychology, 29*, 232-244. doi: 10.1207/S15374424jccp2902_9

Lewis, A., & Lindsay, G. (2000). *Emerging issues. Researching children's perspectives*, Buckingham, Philadelphia: Open University Press.

LoCasale-Crouch, J., Mashburn, A. J., Downer, J. T., & Pianta, R. C. (2008). Pre-kindergarten teachers' use of transition practices and children's adjustment to kindergarten. *Early Childhood Research Quarterly, 23,* 124-139. doi: 10.1016/j.ecresq.2007.06.001

Maguire, M. C., & Dunn, J. (1997). Friendships in early childhood, and social understanding. *International Journal of Behavioral Development, 21*(4), 669-686. Doi: 10.1080/016502597384613

Mantzicopoulos, P., French, B. F., & Maller, S. J. (2004). Factor structure of the Pictorial Scale of Perceived Competence and Social Acceptance. With two pre-elementary samples. *Child Development*, *75*(4), 1214-1228. doi: 10.1111/j.1467-8624.2004.00734.x

Merrell, K. W. (1994). Preschool and kindergarten behavior scales. Test manual. Brandon: Clinical Psychology Publishing Company.

Merrell, K. W. (2003). *Preschool and kindergarten behavior scales. Examiner's manual.* (2nd ed.). Pro-ed.and International Publesher.

Morris, A. S., Silk, J. S., Steinberg, L., Sessa, F. M., Avenevoli, S., & Essex, M. J. (2002). Temperamental vulnerability and negative parenting as interacting predictors of child adjustment. *Journal of Marriage and Family*, *64*, 461-471. doi: 10.1111/j.1741-3737.2002.00461.x

Morrison, F. J., Ponitz, C. C., & McClelland, M. M. (2010). Self-regulation and academic achievement in the transition to school. In S. D. Calkins & M. A. Bell (Eds.), *Human brain development. child development at the intersection of emotion and cognition* (pp. 203-224). Washington, DC, US: American Psychological Association.

Murray, C., Murray, K. M., & Waas, G. A. (2008). Child and teacher reports of teacherstudent relationships: Concordance of perspectives and associations with school adjustment in urban kindergarten classrooms. *Journal of Applied Developmental Psychology, 29,* 49-61. doi: 10.1016/j.appdev.2007.10.006

Myers, S. S. (2007). *Contextual and dispositional influences on low-income children's school adjustment.* (Doctoral dissertation). Retrieved from https://scholarworks.uno.edu/cgi/viewcontent.cgi?article=1558&context=td

Nur, İ., Aktaş-Arnas, Y., Abbak, B. S., & Kale, M. (2018). Mother-child and teacher-child relationships and their associations with school adjustment in pre-school. *Educational Sciences: Theory & Practice, 18,* 201–220. doi: 10.12738/estp.2018.1.0608

O'Connor, E., & McCartney, K. (2007). Examining teacher-child relationships and achievement as part of an ecological model of development. *American EducationalResearch Journal, 44*(2), 340-369. doi: 10.3102/0002831207302172

Olson, S. L., & Rosenblum, K. (1998). Preschool antecedents of internalizing problems in children. *Early Education and Development, 9,* 117 - 129. doi: 10.1207/s15566935eed0902_1

Önder, A., & Gülay, H. (2010). 5-6 yaş çocukları için okula uyum öğretmen değerlendirme ölçeğinin güvenirlik ve geçerlik çalışması [Reliability and Validity of the Teacher Rating Scale of School Adjustment for 5-6 Years of Children]. *International Online Journal of Educational Sciences, 1*(2), 204-224.

Özbey, S. (2009). *Study of the validity and reliabilty of the Preschool and Kindergarten Behaviour Scales (PKBS-2) and to examine affect of the promoter education program*. (Unpublished doctoral dissertation). Gazi University, Ankara, Turkey.

Pears, K. C., Kim, H. K., Capaldi, D. K., David, C. R., & Fisher P. A. (2012). Father–child transmission of school adjustment: A prospective inter generational study. *Developmental Psychology 49*, 792-803. doi: 10.1037/a0028543

Pianta, R., La Paro, K., Payne, C., Cox, M., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *Elementary School Journal*, *102*(3), 225-238. doi: 10.1086/499701

Pianta, R. C., & McCoy, S. J. (1997). The first day of school: The predictive validity of early school screening. *Journal of Applied Developmental Psychology*, *18*, 1–22. doi: 10.1016/S0193-3973(97)90011-3

Pianta, R. C., & Steinberg, M. S. (1992). Teacher-child relationships and the process of adjusting to school. In R. C. Pianta (Ed.), *New directions for child development, No. 57. Beyond the parent: The role of other adults in children'slives* (pp. 61-80). San Francisco, CA, US: Jossry-Bass Inc.

Pianta, R. C., & Stuhlman, M. W. (2004). Teacher-child relationships and children’s success in the first years of school. *School Psychology Review, 33*(3), 444-458.

Reynolds, A. J. (1991). Early schooling of children at risk. *American Educational Research Journal*, *28*(2), 392-422. doi: 10.2307/1162946

Reynolds-Keefer, L., Johnson, R., Dickenson, T., & McFadden, L. (2009). Validity issues in the use of pictorial Likert scales. *Studies in Learning, Evaluation Innovationand Development, 6,* 15-24. doi: 10.1177/00131649921969811

Rimm-Kaufman, S. E., Pianta, R. C., & Cox, M. J. (2000). Teachers’ judgments of problems in the transition to kindergarten. *Early Childhood Research Quarterly, 15,* 147-166. doi: 10.1016/S0885-2006(00)00049-1

Robinson, C. D. (2013). *Associations among preschool classroom climate, children's social interpersonal skills, and early elementary school adjustment in children at-risk for school failure.* (Unpublished doctoral dissertation**)**. Purdue University, West Lafayette, Indiana.

Sadık, F. (2002). Disiplin yaklaşımları [Disciplinary Approaches]. In Y. Aktaş Arnas, & F. Sadık (Eds.). *Okul öncesinde sınıf yönetimi* [Classroom management in pre-school] (pp.163-197). Ankara: Pegem Yayınevi.

Saft, E. W., & Pianta, R. C. (2001). Teachers' perceptions of their relationships with students: Effects of child age, gender, and ethnicity of teachers and children. *School Psychology Quarterly, 16,* 125-141. doi: 10.1521/scpq.16.2.125.18698

Sasser, T. R., Bierman, K. L., & Heinrichs, B. (2015). Executive functioning and school adjustment: The mediational role of pre-kindergarten learning-related behaviors. *Early Childhood Research Quarterly, 30,* 70–79. doi: 10.1016/j.ecresq.2014.09.001

Schulting, A. B., Malone, P. S., & Dodge, K. A. (2005). The effect of school-based kindergarten transition policies and practices on child academic outcomes. *Developmental Psychology*, *41*(6), 860-871. doi: 10.1037/0012-1649.41.6.860

Sette, S., Hipson, W. E., Zava, F., Baumgartner, E., baiocco, R., & Coplan, R. J. (2018). Linking shynss with social and school adjustment in early childhood: The moderating role of inhibitory control. Early Education & Development [Special Issue-Moving Forward in the Study of Temperament and Early Education Outcomes: Mediating and Mederating Factors], 5, 675-690. doi: 10.1080/10409289.2017.1422230

Sette, S., Hipson, W. E., Zava, F., Baumgartner, E., & Coplan, R. J. (2018). Linking shyness with social and school adjust-ment in early childhood: The moderating role of inhibitory control. *Early Education and Development, 29*, 675–690.

Smith, J. (2011). Measuring School Engagement: A Longitudinal Evaluation of the School Liking and Avoidance Questionnaire from Kindergarten through Sixth Grade. (Master's thesis). Retrieved from https://repository.asu.edu/items/9309

Tomada, G., Schneider, B., de Domini, P., Greenman, P., & Fonzi, A. (2005). Friendship as a predictor of adjustment following a transition to formal academic instruction and evaluation. *International Journal of Behavioral Development, 29*(3), 14-322. doi: 10.1177/01650250544000099

Vassilopoulos, S. P., Brouzos, A., & Koutsianou, A. (2018). Outcomes of a universal social and emotional learning (SEL) group for facilitating first-grade students' school adjustment. *International Journal of School & Educational Psychology, 6*(3), 1-14. doi: 10.1080/21683603.2017.1327830

Williams, K. E., Nicholson, J. M., Walker, S., & Berthelsen, D. (2016). Early childhood profiles of sleep problems and self-regulation predict later school adjustment. *British Journal of Educational Psychology, 86*(2), 331-50. doi: 10.1111/bjep.12109

Yoleri, S. (2014). The effects of age, gender, and temperament traits on school adjustment for preschool children. *International Journal of Educational Research, 5*(2), 54-66. doi: 10.19160/e-ijer.55208

Yoleri, S. (2015). Preschool children's school adjustment: indicators of behaviour problems, gender, and peer victimisation. *Education 3 - 13, 43*(6), 630 - 640. doi: 10.1080/03004279.2013.848915

Zhang, L. (2016). *Social and school-related correlates of shyness and unsociability in Chinese adolescents.* (Doctoral dissertation). Retrieved from https://repository.asu.edu/attachments/176487/content/Zhang_asu_0010E_16482.pdf

Zhang, J. J., Smith, D. W., Lam, E. T. C., Brimer, J., & Rodriquez, A. (2002). Development of an evaluation scale to measure participant perceptions of after-school enrichment programs. *Measurement in Physical Education and Exercise Science*, 6, 167-186. doi: 10.1207/S15327841MPEE0603_2

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 555–567

https://dx.doi.org/10.21449/ijate.634091

Published at http://www.ijate.net http://submit.ijate.net/en Research Article

Analyzing the Maximum Likelihood Score Estimation Method with Fences in ca-MST

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

[1] Gazi University, Gazi Education Faculty, Department of Educational Sciences, Turkey
[2] Hacettepe University, Lifelong Learning Center, Turkey

ARTICLE HISTORY

Received: 30 July 2019

Revised: 01 October 2019

Accepted: 16 October 2019

KEYWORDS

MLEF,
MLE,
EAP,
ca-MST,
Ability estimation

Abstract: New statistical methods are being added to the literature as a result of scientific developments each and every day. This study aims at investigating one of these, Maximum Likelihood Score Estimation with Fences (MLEF) method, in ca-MST. The results obtained from this study will contribute to both national and international literature since there is no such study on the applicability of MLEF method in ca-MST. In line with the aim of this study, 48 conditions (4 module lengths (5-10-15-20) x 2 panel designs (1-3; 1-3-3) x 2 ability distribution (normal-uniform) x 3 ability estimation methods (MLEF-MLE-EAP) were simulated and the data obtained from the simulation were interpreted with correlation, RMSE and AAD as an implication of measurement precision; and with conditional bias calculation in order to show the changes in each ability level. This study is a post-hoc simulation study using the data from TIMSS 2015 at the 8th grade in mathematics. "xxIRT" R package program and MSTGen simulation software tool were used in the study. As a result, it can be said that MLEF, as a new ability estimation method, is superior to MLE method in all conditions. EAP estimation method gives the best results in terms of the measurement precision based on correlation, RMSE and AAD values, whereas the results gained via MLEF estimation method are pretty close to those in EAP estimation method. MLE proves to be less biased in ability estimation, especially in extreme ability levels, when compared to EAP ability estimation method.

1. INTRODUCTION

Individualized tests have been administered together with computer technology for a long time. These tests, also known as Computer Adaptive Tests (CAT), are using Item Response Theory in the background. The relationship between IRT, latent ability and item parameter is continuous and defined with monotonic mathematical function (Embretson & Raise, 2000; Reckase, 2009). In this way, the test administration algorithm is designed so that the test items which are administered to the test taker are adapted in terms of difficulty in line with the test taker's estimated ability while the test is going on. As the individuals receive items appropriate

CONTACT: Melek Gülşah ŞAHİN ✉ mgulsahsahin@gazi.edu.tr Gazi University, Gazi Education Faculty, Department of Educational Sciences, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

to their own level of ability, they do not get the same test form as is the case in pen-and-paper tests. Also, it is prevented that the individuals receive items which are way above or under their estimated ability. A CAT is more effective than a regular test by having the appropriate item pool (Wainer, Kaplan, & Lewis 1992). Using computer technology has provided the users with convenient strategies such as administering and securing the test items as well as analyzing and storing the data easily. Because of the aforementioned virtues, CAT ensures more efficient and precise measurement in individuals' ability distribution. Although CAT has gained a sound ground in terms of application in a variety of fields, it has its own limitations. Some of them can be listed as being difficult to apply in different item formats, requiring a large item pool as well as complicated software and fast computers, not enabling test items to be revised throughout the test, having a complex item selection algorithm, and not being able to get information about psychometric characteristics since test formats are established during the test (Hambleton, Swaminathan, & Rogers, 1991; Hendrickson, 2007; Luecht & Nungester, 1998; Luecht & Sireci, 2011; Sarı, Yahşi Sarı, & Huggins Manley, 2016; Yan, von Davier, & Lewis, 2014).

Due to its limitations, CAT is gradually being replaced by Computer Adaptive Multistage Tests (ca-MST). ca-MST combines the characteristics of linear and adaptive tests. While the appropriate models are chosen according to the individuals' level in the application as is the case in adaptive tests, the test takers can revise the test items as they can do in linear tests and test content is generally set before the test is administered (Leucht & Nungester, 1998). In these tests, there is an adaptation, not on an item basis, but on the basis of item sets called modules (Leucht & Nungester, 1998; Yan, von Davier & Lewis, 2014; Zenisky, Hambleton & Leucht, 2010).

ca-MST is more advantegous than CAT especially because of the fact that the test formats in ca-MST can be examined in advance by test developers and test items can be reviewed by test takers during test administration (Luecht, Brumfield, & Breithaupt, 2006; Hendrickson, 2007).

1.1. Ability Estimation Methods

In individualized tests, which items will be set to a test taker is not decided beforehand. It is necessary to estimate the individual's ability to be able to choose the items. Based on the individuals' performance, the next item which is appropriate to the individual's ability is chosen from the pool which has specific item parameters. Different from CAT, in ca-MST, the individual's ability is estimated after each module and the most appropriate module in the first stage that comes after the estimation is administered to the individual.

Since IRT-based estimation methods are used in estimating the individual's ability, the ones used for CAT can also be used for ca-MST. In the literature, the most frequently used ability estimation methods are Maximum Likelihood Estimation (MLE), Expected a posteriori (EAP), and Maximum a posteriori (MAP) (Baker & Kim, 2004; Embretson & Resie, 2000). MLE method is often chosen because it is based on likelihood function and it provides unbiased estimates. The log likelihood function of an individual which was estimated after an administered test item is represented below.

$$L = \ln(\mu|\theta) = \sum_{j=1}^{n}\left[\mu_j \ln P_j + \left(1 - \mu_j\right)\ln\left(1 - P_j\right)\right]$$

where μ is a response string of j items, which is (μ1, μ2, μ3..), and Pj is the item response function given theta (θ).

In MLE method, the module that provides the most information about the individual is chosen. Although the ML estimator is efficient and unbiased in asymptotical terms, a large item pool is needed to make use of it while it is not applicable with examinees having all-endorsed or all-not-endorsed response patterns (Embretson & Reise, 2000). Therefore, this method requires individuals to have at least one correct and one incorrect response in order to estimate abilities.

In response to the limitations of MLE, Bayesian-based estimation methods are proposed. These suggested methods include Modal a posteriori – MAP (Samejima, 1969) and Expected a posteriori –EAP (Bock & Mislevy, 1982). The MAP estimator combines the available information in hand and exclusive trait level true for all kinds of possible response patterns. What is problematic about MAP is that it might give biased results when the tests are short (e.g., <20), especially when the prior is used in an incorrect way (Embretson & Reise, 2000).

Contrary to the ML and MAP estimators, EAP estimation of trait level requires a non-iterative process. Unlike ML estimation, EAP yields a finite trait level estimation for all response patterns, including endorsed and not-endorsed ones (Embretson & Reise, 2000). If the item number is finite, the EAP estimator ise biased. The type of bias can be described as that the trait level is biased when the item number is finite (Wainer & Thissen, 1987). In EAP and MAP estimation methods, the item is selected in a way to decrease the individual's ability estimation range to minimum, and ability estimation is done in all kinds of response patterns. Although EAP and MAP estimation methods are similar, there are some significant differences between them. EAP estimation requires a discrete prior contrary to a continuous prior. Because of that, EAP is a scoring strategy that is used most easily among IRT models and testing context (Embretson & Reise, 2000).

In literature, the methods different from MLE and Bayesian-based methods are examined especially for bias reduction (Firth,1993; Magis & Raiche, 2010; Magis, Beland & Raiche, 2010). In summary, each ability estimation method has its own limitations. Han (2016) has developed a method called maximum likelihood estimation with fences (MLEF) to eliminate those method's limitations. Although this method is basically similar to MLE, it requires for score estimation that the MLEF places two imaginary items having fixed responses in order to build ''fences'' around a meaningful range of the log likelihood function. In MLEF, the first imaginary item is accepted to be the lower fence and its b parameter is set at theta, where the lower bound of the theta distribution is expected (e.g., b = -3.5). For the b parameter value, the lower fence should not be higher than any other item included in the test form. Similarly, the second imaginary item is accepted to be the upper fence, and its b parameter is set at u, where the upper log likelihood functions of three-item response patterns bound of the theta distribution is expected (e.g., b = 3.5). The b-parameter upper fence value should be larger than any other item included in the test form. These two ''fence'' items should be established to possess a very high a-parameter value (e.g., a = 3.0). The log likelihood function estimated in MLEF method is presented below.

$$L^* = \ln P_{LF} + \ln(1 - P_{UF}) + \sum_{j=1}^{n} \left[\mu_j \ln P_j + \left(1 - \mu_j\right) \ln\left(1 - P_j\right)\right]$$

where P_{LF} and P_{UF} are the item response functions of the lower and upper fences.

1.2. Purpose of the Study

Bayesian-based EAP and MAP methods, which are suggested to eliminate some limitations of MLE, one of the most frequently used methods in literature, are known to result in estimates toward the center of a prior distribution, resulting in a shrunken score scale (Weiss and McBride, 1984). There are limited studies about ability estimation methods in ca-MST in literature. One of them is the study carried out by Kim, Moses and Yoo (2015). In that study, researchers have compared MLE, EAP, MAP ve TCF (test characteristic function) methods with different grading methods for tests having different module length. The study is important as the method, which is previously examined only in CAT, is being examined in ca-MST in different conditions; it provides a comparison of the method with other frequently-used methods in literature; and there is no similar study in the literature. Besides, what makes this study so important is that it compares the method presented by Han (2016) with other existing methods.

The aim of this study is to investigate the effect of MLEF that is developed to eliminate limitations of the related methods on ability estimation in ca-MST. And also, the applicability of MLEF method for ca-MST and the comparison of MLE method that is often used and referred to in literature and Bayesian-based EAP methods for different test conditions are investigated.

2. METHOD

In this study, it is aimed to investigate the effect of different ability estimation methods on ability estimation in ca-MST. For that purpose, real item data were used in the study. Therefore, this study is a descriptive research based on post hoc simulation that uses real item parameters.

2.1. Obtaining of Item Parameters

In the study, an item pool made from TIMSS 2015 mathematics-eight grade assessment items was used. Two item formats are used in the TIMSS assessments: multiple-choice and constructed response. Multiple-choice items represent at least half of the total number of points. Items are grouped into a series of item blocks in TIMSS assessment. Approximately 12–18 items in each block at the eigth grade and a total of 28 blocks were assigned to 14 different achievement booklets at each grade level in TIMSS 2015 (IEA, 2013).

In TIMSS 2015, there are 297 eight grade mathematics items in total. Within the scope of this study, parameters of 159 items in total, which are estimated based on 3-parameter logistics model (Lord, 1980) and graded based on 1-0, are taken from the website https://timssandpirls.bc.edu/timss2015/international-database/. Table 1 shows the descriptive characteristics of parameters belonging to the items handled in the study. Within the scope of the study, an MST item pool is formed with 159 items in total.

Table 1. Mean and standard deviation of item parameters

Parameters	Mean	SD
a	1.31	0.39
b	0.51	0.53
c	0.21	0.08

2.2. Simulee Parameters

In this study, two different distribution types; namely, normal distribution N (0,1) and uniform distribution (-3,3), are examined. Two distribution types are chosen in order to be able to compare MLEF methods with others in case the numbers of simulees, especially at peak ability levels, are different. Therefore, 5000 simulee parameters having both normal and uniform distribution are simulated by using MSTGen (Han, 2013) simulation software tool.

2.3. ca-MST Components

Within the scope of this study, 1-3 (Patsula, 1999; Kim, Moses, & Yoo, 2015) and 1-3-3 (Jodoin, Zenisky, & Hambleton, 2006; Leucht, Brumfield, & Breithaupt; 2006; Park, 2015; Patsula, 1999; Zenisky, 2004) panel designs, which are frequently used in the literature, were studied as in one panel. TIF values used in forming the panels are specified as below in Figure 1.

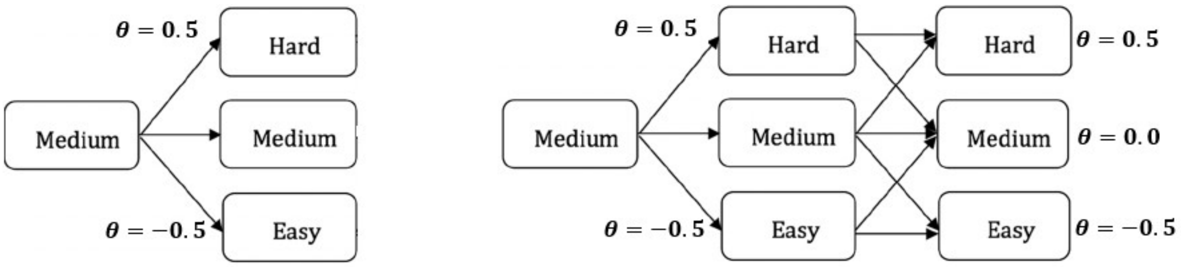


Figure 1. TIF values of 1-3, and 1-3-3 panel designs

In this study, four different module lengths (5-10-15-20) are examined because module lengths vary from small (5 to 10) to large (50 to 100 items) (Luecht, 2000). The lengths of the modules used within the scope of this study are different because test length in ca-MST is correlated with measurement precision (Patsula, 1999), and this study aims to display the effects of ability estimation methods more clearly in long tests.

MLE, EAP and MLEF methods are used for ability estimation of simulees. Maximum Fisher Information is used as item selection method, and "bottom up" is used as test assembly method. For test assembly process, "xxIRT" (Luo, 2017) package program is used in R software (R Development Core Team, 2011).

Table 2 shows 48 conditions examined in this study (2 ability distribution ×2 panel design × 4 module length × 3 ability estimation method).

Table 2. ca-MST components

Components	Variables
Examinee distribution	Normal-Uniform
Panel Design	"1-3"; "1-3-3"
Module Length	5-10-15-20
Estimation method	MLE-EAP-MLEF

2.4. Data Analysis

After MST conditions are created for each variable specified in Table 2, simulee parameters and test conditions are matched up within the context of conditions specified in the study with MSTGen software.

In this study, for each condition, correlation (between the simulated / derived thetas and estimated thetas calculating after ca-MST) root mean square error (RMSE), and average absolute difference (AAD) values are calculated. Pearson's Product Moments Correlation is used in calculating the correlation coefficient. And also, the equations of RMSE and AAD are presented below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}},\ AAD = \frac{\sum_{i=1}^{n}|\hat{\theta}_i - \theta_i|}{n}$$

where $\hat{\theta}_i$ represents the estimated level of ability for person i, θ_i represents the known level of ability for person i, and n represents the size of the sample.

In addition to those, it is aimed to examine the changes in ability levels based on bias values in detail. With this aim, ability levels are grouped based on changes of theta 0.5; and bias values are examined in 12 θ change points in uniform distribution and in 15 θ change points in normal distribution (Zenisky, 2004).

3. FINDINGS

In this section, data gathered from the study are presented in two parts. Correlation, RMSE, and AAD values are given in the first part; and conditional bias values are given in the second one.

3.1. Results of Correlation, RMSE and AAD

Correlation, RMSE and AAD values of 48 conditions examined in the study are presented in Table 3. When the correlation values in Table 3 are examined, it is seen that correlation values generally increase when the panel design shifts from two-stage structure to a three-stage one. In panel design 1-3, the highest correlation value of 0.9679 is obtained under the condition when the module length is composed of 20 items and the items are administered to simulees within the uniform ability distribution having EAP ability level estimation method. On the other hand, the lowest correlation value of 0.6491 is obtained under the condition when the module length is composed of 10 items and the items are administered to simulees in normal ability distribution having MLE ability level estimation method. In panel design 1-3-3, the highest correlation value of 0.9770 is obtained under the condition when the module length is composed of 20 items and the items are administered to simulees within the uniform ability distribution having EAP ability level estimation method. On the other hand, the lowest correlation value of 0.6614 is obtained under the condition when the module length is composed of 5 items and the items are administered to simulees within the normal ability distribution having MLE ability level estimation method.

When ability level estimation methods are examined, it is seen that the highest correlation values are obtained in EAP ability estimation method. This is valid for both two-stage and three-stage panel designs. It is also valid when the number of items in modules increases. However, in MLE method, under the condition when module length is composed of 5 items, correlation value is higher than the results under the conditions when module is longer.

It is seen that there is a general increase in correlation values as the number of items in modules increases. Moreover, correlation values in normal ability distribution conditions are lower compared to the ones in uniform ability distribution conditions.

When RMSE values are examined, it is seen that RMSE values generally decrease when the panel design shifts from two-stage structure to a three-stage one. In panel design 1-3, the highest RMSE value of 6.0165 is obtained under the condition when the module length is composed of 5 items and the items are administered to simulees within the uniform ability distribution having MLE ability level estimation method. On the other hand, the lowest RMSE value of 0.2949 is obtained under the condition when the module length is composed of 20 items and the items are administered to simulees within the normal ability distribution having EAP ability level estimation method. In panel design 1-3-3, the highest RMSE value of 4.2494 is obtained under the condition when the module length is composed of 5 items and the items are administered to simulees within the uniform ability distribution having MLE ability level estimation method. On the other hand, the lowest RMSE value of 0.2515 is obtained under the condition when the module length is composed of 20 items and the items are administered to simulees within the normal ability distribution having EAP ability level estimation method.

When ability estimation methods are examined, it is seen that lower RMSE values are obtained in EAP ability estimation method. This is valid for both two-stage and three-stage panel designs. It is also valid when the number of items in modules increases. RMSE values decrease as the number of items in modules increase. Moreover, RMSE values in normal ability distribution conditions are lower compared to the ones in uniform ability distribution conditions. RMSE values obtained via MLEF ability estimation method are found to be a bit higher than the values obtained via EAP estimation method at the two stage panel design as well as the three stage panel design. When compared to MLE, RMSE values obtained via MLEF ability estimation method can be said to be quite low. Moreover, the lowest RMSE value is obtained at normal distribution at both panel design and all module lengths when MLEF method is adopted.

When AAD values are examined, it is seen that AAD values generally decrease when the panel design shifts from two-stage structure to a three-stage one. In panel design 1-3, the highest AAD value of 4.5277 is obtained under the condition when the module length is composed of 5 items and the items are administered to simulees within the uniform ability distribution having MLE ability level estimation method. On the other hand, the lowest AAD value of 0.2133 is obtained under the condition when the module length is composed of 20 items and the items are administered to simulees within the normal ability distribution having EAP ability level estimation method. In panel design 1-3-3, the highest AAD value of 2.4339 is obtained under the condition when the module length is composed of 5 items and the items are administered to simulees within the uniform ability distribution having MLE ability level estimation method. On the other hand, the lowest AAD value of 0.1812 is obtained under the condition when the module length is composed of 20 items and the items are administered to simulees within the normal ability distribution having EAP ability level estimation method.

When ability estimation methods are examined, it is seen that the lowest AAD values are obtained in EAP ability estimation method. This is valid for both two-stage and three-stage panel designs. It is also valid when the number of items in modules increases. However, the values obtained in both MLEF and EAP ability estimation methods are very close.

AAD values generally increase as the number of items in modules increase. Moreover, AAD values in normal ability distribution conditions are lower compared to the ones in uniform ability distribution conditions.

3.2. Results of Conditional Bias

Conditional bias values calculated in groups according to the changes of 0.5 in ability levels are given in Figure 2. When it is examined, it is seen that under conditions when the test is administered to simulees in normal ability distribution, bias values are higher in extreme ability levels independently of panel design, test length and ability estimation methods. However, the highest bias values in extreme ability levels are under conditions when MLE estimation method is used. While bias values approach the negative infinity as the move is towards -3 ability level, bias values approach the positive infinity as the move is towards +3 ability level. Furthermore, under conditions when two-stage panel design and MLE ability estimation methods are used, more errors are obtained when compared to other methods in mid-levels especially under the condition when the module length is composed of five items. When EAP and MLEF methods are compared independently from stage number, lower bias values are gathered in especially negative extreme points of MLEF method compared to EAP method.

Table 3. Correlation, RMSE and AAD results of ability estimation

PD	AD	AEM	CORRELATION				RMSE				AAD			
			ML5	ML10	ML15	ML20	ML5	ML10	ML15	ML20	ML5	ML10	ML15	ML20
PD1-3	UNIFORM	MLE	0.8517	0.7560	0.7770	0.7655	6.0165	4.3157	4.3010	3.7308	4.5277	2.4887	2.4306	1.9440
PD1-3	UNIFORM	EAP	0.9381	0.9539	0.9645	0.9679	0.7236	0.6226	0.5415	0.5123	0.5453	0.4512	0.3869	0.3623
PD1-3	UNIFORM	MLEF	0.9124	0.9338	0.9501	0.9556	0.7508	0.6377	0.5548	0.5234	0.5421	0.4519	0.3903	0.3654
PD1-3	NORMAL	MLE	0.7518	0.6491	0.6672	0.6664	4.6654	3.3977	3.2046	2.6278	2.6395	1.4965	1.3239	0.9748
PD1-3	NORMAL	EAP	0.8998	0.9321	0.9492	0.9571	0.4438	0.3684	0.3202	0.2949	0.3367	0.2729	0.2356	0.2133
PD1-3	NORMAL	MLEF	0.8621	0.9026	0.9277	0.9381	0.6591	0.5185	0.4345	0.3975	0.4663	0.3510	0.2886	0.2588
PD1-3-3	UNIFORM	MLE	0.7610	0.7591	0.7558	0.7759	4.2494	3.3880	3.2458	2.5211	2.4339	1.7085	1.5556	1.0685
PD1-3-3	UNIFORM	EAP	0.9510	0.9657	0.9709	0.9770	0.6401	0.5327	0.4854	0.4254	0.4711	0.3807	0.3424	0.2966
PD1-3-3	UNIFORM	MLEF	0.9314	0.9516	0.9592	0.9689	0.6507	0.5468	0.5000	0.4365	0.4685	0.3842	0.3464	0.3008
PD1-3-3	NORMAL	MLE	0.6614	0.6832	0.6637	0.7480	3.0762	2.1232	2.2318	1.3407	1.2925	0.7483	0.7713	0.4056
PD1-3-3	NORMAL	EAP	0.9248	0.9515	0.9609	0.9690	0.3871	0.3130	0.2819	0.2515	0.2904	0.2295	0.2038	0.1812
PD1-3-3	NORMAL	MLEF	0.8959	0.9303	0.9420	0.9572	0.5396	0.4242	0.3847	0.3190	0.3741	0.2804	0.2480	0.2087

PD: Panel Desing, AD: Ability Distribution, AEM: Ability Estimation Method, ML: Module Length

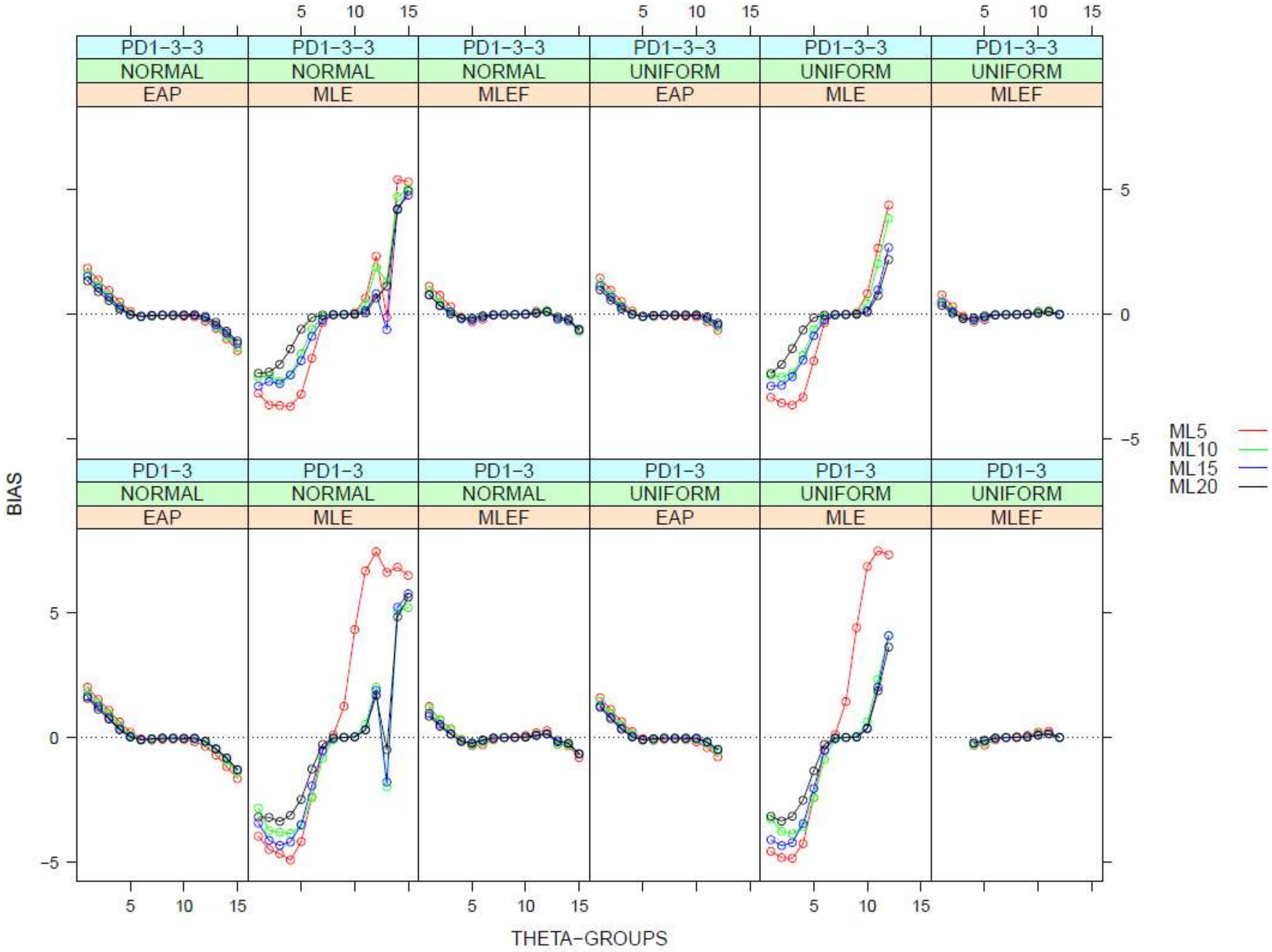


Figure 2. Conditional bias in ability estimation

Similar to the condition when the examinees have normal distribution, under conditions when the items are administered to examinees having uniform ability distribution, the bias values in extreme ability levels are obtained mostly in the one where MLE estimation method is used. Again, the lowest bias value in this condition is obtained with MLEF method.

As the distribution of individuals changes from normal towards uniform, lower bias values are obtained under the condition when the examinees in uniform ability distribution take the test in each ability estimation method. In other words, a good level of estimation can be done under conditions when MLEF method is used.

In general, as the module length changes, the change in bias values can be seen more clearly in MLE method. Especially the bias values obtained in 5-item module length are more noteworthy than the ones in other module lengths. When different module lengths are examined in EAP and MLEF methods, there is a slight change in bias values under conditions when the items are administered to examinees having normal distribution compared to the change in module length especially in extreme points of EAP ability estimation method. However, this situation is less obvious under conditions when MLEF method is used. When the ability distribution is uniform, MLEF ability estimation method has lower bias values in extreme ability levels compared to EAP.

When the graphs obtained from the conditions of different stage numbers are examined, it is seen that there are no significant differences in errors obtained in conditions where EAP and MLEF methods are used. However, lower values are obtained in extreme ability levels of panel design 1-3-3 where MLE method is used. Besides, bias values in mid-ability levels of both uniform and normal ability distributions, where three-stage condition is examined, are in a wider range, around 0.

4. DISCUSSION and CONCLUSION

In this study, the aim is to investigate what results MLEF ability level estimation method, which is brought to literature by Han (2016), gives in terms of ca-MST ability estimation when compared to MLE and EAP methods. In line with this, 48 conditions in total are examined with different panel designs, module lengths and individuals who have different ability level distributions. When the data are interpreted, it is seen that generally MLEF method is more successful in both short and long tests compared to MLE method; and it is successful in decreasing bias values especially in extreme ability levels compared to EAP method.

When correlation, RMSE and AAD values are examined in the study as the indicators of measurement precision, the precision is lower under conditions when MLE ability estimation method is used. Although this result changes as the number of items or stages in modules increases, the results are not close to the values gathered in MLEF or EAP ability estimation methods. It can be said that the result is an expected one considering that there needs to be at least one correct and one incorrect answer in order for MLE estimation method to conduct ability estimation. When the measurement precision values of EAP and MLEF ability estimation methods are examined, it is seen that the values are very close to each other, but EAP method provides a more precise measurement. However, another result is that the differences of correlation, RMSE and AAD values in both methods decrease as the number of items in modules increases. The results are valid for both normal and uniform distribution. When different conditions in two- and three-stage conditions are compared, measurement precision is higher in three-stage conditions. This can be explained by the fact that there is one adaptation point in two-stage tests; in other words, by the fact that there are less measurement results in estimating the simulee's ability level.

When the data gathered in the study are examined in terms of conditional bias values based on ability levels, it is seen that MLE ability estimation method is extremely biased, especially in extreme ability levels. The bias values reach the maximum level, especially in modules where the number of items is five. As the number of items and stages in modules increase, there are slight decreases in bias values. Yen and Fitzpatrick (2006) state in their study that ability estimation whose measurement precision is high can be obtained with MLE method, especially in conditions when the modules are composed of 30 or more items. Besides, it is concluded that it is critical that there are five items in modules for bias values obtained by MLE ability estimation method.

When MLE and EAP ability estimation methods are compared, EAP method shows less bias, especially in extreme ability levels. This result is expected when considering that Bayesian-based methods evolved as a solution to the estimation issues for individuals whose responses are all correct or incorrect in MLE ability estimation method. Similar to the result of this study, Kim, Moses and Yoo (2015) claim that in their two-stage MST study, measurement precision is higher compared to MLE, one of Bayesian-based estimation methods. Also, it is stated that Bayesian-based methods in which MLE is less precise are a better option for high-performing examinees.

When EAP and MLEF ability estimation methods are examined in terms of conditional bias, it is seen that MLEF method has extremely low bias values, especially in extreme ability levels. Therefore, it can be claimed that MLEF method will be slightly biased in estimating abilities, especially of those individuals who have extreme ability levels. Especially when an ability estimation is conducted with ca-MST application for a group whose ability distribution is uniform, bias values are almost around 0 in each ability distribution level. These results are valid even for modules which have the lowest number of items. When Han (2016) compares MLEF method with other estimation methods for CAT in the study, it is stated that similar to this study's results, the estimation can be done with very small bias in extreme ability levels. In line with the results of the study, it is suggested that MLEF method can be preferred over EAP method in terms of providing less biased results. However, especially under conditions when module length is short, it is suggested that test developers can use EAP and MLEF methods for ability estimation instead of MLE method. When deciding on the panel design, since there are more estimation points in three-stage designs, those can be suggested instead of two-stage ones in terms of providing a more measurement precision.

Considering the conditions examined in the study, researchers can further investigate the following issues: trying similar conditions in different panel designs such as 1-2-4; 1-2-2; 1-5-5; 1-2-3-4 where the numbers of items and stages can be changed; examining different ca-MST components such as content balancing or item exposure control; examining similar conditions in different item pools with different item selection methods and examining the effect of different routing module methods on ability level estimation.

ORCID

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

5. REFERENCES

Baker, F.B., & Kim, S. (2004). *The basics of item response theory using R.* New York: Marcel Dekker.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in microcomputer environment. *Applied Pyschological Measurement*, *6*, 431-444. DOI: 10.1177/014662168200600405

Embretson, S. E., and Reise, S.P. (2000). *Item response theory for pyschologists*. Mahwah, NJ, US: Lawrence Erlbaum

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38.

Magis, D., Beland, S., & Raiche, G. (2010). A test-length correction to the estimation of extreme proficiency levels. *Applied Psychological Measurement, 35*, 91–109.

Magis, D., & Raiche, G. (2010). An iterative maximum a posteriori estimation of proficiency level to detect multiple local likelihood maxima. *Applied Psychological Measurement*, *34*, 75–90.

Han, K. T. (2013). MSTGen: simulated data generator for multistage testing. *Applied Psychological Measurement*, *37*(8) 666–668. doi: 10.1177/0146621613499639

Han, K. T. (2016). Maximum Likelihood Score Estimation Method with Fences for Short Length Tests an Computerized Adaptive Tests. *Applied Psychological Measurement*, *40*(4), 289-301.

Hambleton, R. K., H. Swaminathan and H. J. Rogers. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hendrickson, A. 2007. An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*, 44–52.

International Association for the Evaluation of Educational Achievement (IEA), (2013). *TIMSS 2015 Assessment Frameworks*. Boston College: TIMSS & PIRLS International Study Center, Lynch School of Education.

Jodoin, M. G., Zenisky, A. L., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19*(3), 203-220.

Kim, S., Moses, T., & You, H. (2015). A Comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*. 52(1), 70-79.

Luecht, R. M. (2000). *Implementing the Computer-Adaptive Sequantial Testing (CAST) framework to mass produce high quality computer adaptive and mastery tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), New Orleans, LA.

Luecht, R.M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*(3), 189-202.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229-249.

Leucht, R., & Sireci, S.G. (2011). A review of models for computer-based testing. Research Report. New York: The College Board. Retrieved from https://files.eric.ed.gov/fulltext/ED562580.pdf

Luo, X. (2017). *Package 'xxIRT'.* (Version 2.0.3). Retrieved September 25, 2018 from https://cran.r-project.org/web/packages/xxIRT/xxIRT.pdf

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Park, R. (2015). *Investigating the impact of a mixed-format item pool on optimal test design for multistage testing.* (Unpublished doctoral dissertation). University of Texas at Austin.

Patsula, L. N. (1999). *A comparison of computerized-adaptive testing and multi-stage testing.* (Unpublished doctoral dissertation). University of Massachusetts at Amherst.

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

Reckase, M.D. (2009). *Multidimensional item response theory*. New York: Springer

Robin, F. (1999, March). *Alternative item selection strategies for improving test security and pool usage in computerized adaptive testing.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montréal, Québec.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrica Monograph Supplement, 34*(4,Pt.2), 100)

Sarı, H.İ., Yahşi Sarı, H., & Huggins Manley, A. C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(2), 388-406. DOI: 10.21031/epod.280183

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement*, *29*(3), 243-251.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*, 339-368.

Yan, D., von Davier, A.A., & Lewis, C. (2014) *Computerized multistage testing: Theory and applications*. CRC Press

Yen, W. M., & Fitzpatrick, A.R. (2006). Item response theory. In R. L. Brennan (Ed.). *Educational measurement.* Westport, CT: American Council on Educaiton and Praeger.

Zenisky, A. L. (2004). Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment. (Unpublished doctoral dissertation). University of Massachusetts at Amherst.

Zenisky, A., Hambleton, R.J., & Luecht, R.M. (2010). Multistage testing: Issues, design and research. In W.J. ven der Linden & C.E.W. Glass (Eds.). *Elements of adaptive testing* (pp.355-372). New York: Springer

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 568–579
https://dx.doi.org/10.21449/ijate.581314

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

The Uniform Prior for Bayesian Estimation of Ability in Item Response Theory Models

Tuğba Karadavut [1,*]

[1] Recep Tayyip Erdogan University, Faculty of Education, Cayeli, Rize, Turkey

ARTICLE HISTORY

Received: 22 June 2019
Revised: 16 September 2019
Accepted: 17 October 2019

KEYWORDS

Item response theory,
Uniform ability,
Bayesian estimation

Abstract: Item Response Theory (IRT) models traditionally assume a normal distribution for ability. Although normality is often a reasonable assumption for ability, it is rarely met for observed scores in educational and psychological measurement. Assumptions regarding ability distribution were previously shown to have an effect on IRT parameter estimation. In this study, the normal and uniform distribution prior assumptions for ability were compared for IRT parameter estimation when the actual distribution was either normal or uniform. A simulation study that included a short test with a small sample size and a long test with a large sample size was conducted for this purpose. The results suggested using a uniform distribution prior for ability to achieve more accurate estimates of the ability parameter in the 2PL and 3PL models when the true distribution of ability is not known. For the Rasch model, an explicit pattern that could be used to obtain more accurate item parameter estimates was not found.

1. INTRODUCTION

Item Response Theory (IRT) is widely used in psychological measurement (Embretson, 1996), and in educational measurement (Lord & Novick, 1968) for designing and analyzing the measurement instruments. It also has applications in other fields such as public health, ecology and sociology. In educational measurement, the student ability is the subject of the measurement. Ability is a latent trait and it cannot be measured directly. Thus, student responses to items in a test are used to measure ability in educational measurement. IRT defines a continuous and monotonic mathematical function (Reckase, 2009) for explaining the relationship between latent ability and student responses to the test items (Embretson & Reise, 2000). In this study, the latent ability is assumed to be unidimensional, and the IRT models that are for analyzing the unidimensional latent ability are considered for the analysis.

The estimation methods for IRT models require an assumption regarding the ability distribution to enable estimation of the model parameters. The tradition is to assume a normal distribution for abilityfor estimating the model parameters. Generally, normality is a reasonable assumption for ability (Embretson & Reise, 2000). However, it is not unlikely for observed scores in educational and psychological measurement to be non-normal in reality (e.g., Cook, 1959; Lord, 1955; Micerri, 1989). Micerri (1989), in example, examined 440 raw-score distributions

CONTACT: Tuğba Karadavut ✉ tugba-mat@hotmail.com Recep Tayyip Erdogan University, Faculty of Education, Cayeli, Rize, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

from large-scale achievement and psychometric measures. Micerri (1989) found that, of the measures he investigated, 125 were moderately asymmetric (i.e., 28.4%), and 135 were extremely asymmetric (i.e., 30.7%). The non-normality in the observed scores may also indicate non-normality in the latent ability scores. It is because the observed raw scores and the latent ability scores from an IRT model are correlated (Fan, 1998; Stewart, 2012).

There are two general methods for estimation of the parameters in IRT models. These are marginal maximum likelihood estimation (Bock & Aitkin, 1981) and Bayesian estimation methods. Both of these methods make prior assumptions regarding the ability distribution (Baker & Kim, 2004, de Ayala, 2009). In this study, Markov chain Monte Carlo (MCMC) estimation was used for estimation of the model parameters. MCMC is a Bayesian estimation technique that iteratively samples from the posterior distributions of the parameters to be estimated (Jackman, 2000). These samples are then used to obtain estimates of the parameters. Bayesian estimation methods require indication of a prior distribution for each parameter in the model that is intended to be estimated. The prior distribution for a parameter reflects the distributional assumptions regarding that parameter. Poor specification of the priors in Bayesian estimation may result in biased parameter estimates (e.g., Mislevy, 1986). Therefore, a sufficiently informative prior should be specified for each parameter in the model in order to obtain unbiased estimates of the parameters (Baker & Kim, 2004; Mislevy, 1986). A sufficiently informative prior provides information regarding the posterior distribution of the parameter to be estimated. The prior may be assumed to be from the same distribution family with the posterior distribution (e.g., conjugate prior).

Assumptions with respect to ability distribution have been shown to have an effect on IRT parameter estimation, depending on the deviation from the actual ability distribution (Reise & Yu, 1990; Roberts, Donoghue, & Laughlin, 2002; Sass, Schmitt, & Walker, 2008; Sen, Cohen, & Kim, 2016; Seong, 1990; Stone, 1992). Item parameter estimates are more precise when the prior distribution for latent ability matches the true distribution of latent ability (Seong, 1990). The bias in item parameter estimates due to misspecification of actual ability distribution, on the other hand, often can be reduced by increasing sample size and test length (e.g., de Ayala & Sava-Bolesta, 1999; Kirisci, Hsu, & Yu, 2001, Reise & Yu, 1990; Roberts et al., 2002; Seong, 1990; Stone, 1992). Thus, the effect of the prior distributional assumptions on parameter estimation should be considered with respect to the potentially confounding variables such as the sample size and the test length.

The latent ability distribution in an IRT model can be estimated with an assumption of normality following the general applications in the literature. The true ability distribution, on the other hand, can be in another type such as the uniform distribution (e.g., Hambleton & Cook, 1983; Swaminathan, Hambleton, & Rogers, 2007). In that case, using a prior distribution that matches the true distribution of the latent ability may result in more accurate estimates of item and ability parameters in IRT models. In this study, the normal prior distribution for ability was investigated for its efficiency to result in reasonable estimates of item and ability parameters, especially when the true latent ability distribution was uniform. A simulation study was conducted to analyze student responses to items with a normal and a uniform underlying ability distribution. The analyses were done using a unidimensional IRT model for dichotomous items. The models used in this study were Rasch (Rasch, 1960), two-parameter logistic (2PL; Birnbaum, 1968), and three-parameter logistic (3PL; Birnbaum, 1968) IRT models. Uniform and normal distributions priors were used for the latent ability while analyzing student responses to items. Finally, item and ability parameter estimates from the models with a normal and a uniform prior distribution for the latent ability were compared to the generating item and ability parameters in order to determine the accuracy of item and ability parameter estimates.

2. METHOD

2.1. Unidimensional Item Response Theory Models

Unidimensional IRT models for dichotomous items (e.g., for multiple choice) are extensively used in educational measurement. These models include Rasch, 2PL and 3PL models. The names of 2PL and 3PL models vary depending on the number of item parameters in the model. Namely, the 2PL model has two item parameters that are item difficulty and item discrimination parameters. Similarly, the 3PL model includes three item parameters that are item difficulty, item discrimination and the item pseudo-guessing parameters. The 3PL model defines the probability that an examinee j with ability θ answers item i correctly ($P_i\left(\theta_j|X=1\right)$) with the following equation:

$$P_i\left(\theta_j|X=1\right)=c_i+(1-c_i)\frac{1}{1+e^{-a_i(\theta_j-b_i)}}, \tag{1}$$

where b_i is the item difficulty parameter for item i, a_i is the item discrimination parameter for item i, and c_i is the pseudo-guessing parameter for item i. Fixing the c_i parameter in a 3PL model to zero results the 3PL model to reduce into a 2PL model. Thus, the probability that an examinee j with ability θ answers item i correctly in a 2PL model is:

$$P_i\left(\theta_j|x=1\right)=\frac{1}{1+e^{-a_i(\theta_j-b_i)}}. \tag{2}$$

Similary, fixing the c_i parameter to zero and the $\mathrm{a_i}$ parameter to one in a 3PL model yields a Rasch model. The Rasch model defines the probability that an examinee j with ability θ answers item i correctly as:

$$P_i\left(\theta_j|x=1\right)=\frac{1}{1+e^{-(\theta_j-b_i)}}. \tag{3}$$

2.2. The Simulation Design

Binary student responses to test items were generated using the R (2016) software for the Rasch, 2PL and 3PL models. The underlying latent ability distributions were simulated to follow either a standard normal distribution or a uniform distribution on the interval [-3, 3]. Two test lengths (15-item and 30-item) and two sample sizes (600 and 2,000) were generated. Twenty-five data sets were simulated for each simulation condition. Item parameters that are used to generate student responses to test items are given in Table 1.

Table 1. Item Parameter Estimates Used for Generating Student Responses.

	Rasch	**2PL**		**3PL**		
	b	***b***	***a***	***b***	***a***	***c***
1	2.75	2.75	1.0	2.75	1.0	0.25
2	2.50	2.50	1.0	2.50	1.0	0.25
3	2.25	2.25	1.0	2.25	1.0	0.25
4	2.00	2.00	1.0	2.00	1.0	0.25
5	1.75	1.75	1.0	1.75	1.0	0.25
6	1.50	1.50	1.5	1.50	1.5	0.15
7	1.25	1.25	1.5	1.25	1.5	0.15
8	1.00	1.00	1.5	1.00	1.5	0.15
9	0.75	0.75	1.5	0.75	1.5	0.15
10	0.50	0.50	1.5	0.50	1.5	0.15
11	0.25	0.25	2.0	0.25	2.0	0.10
12	0.00	0.00	2.0	0.00	2.0	0.10
13	-0.25	-0.25	2.0	-0.25	2.0	0.10
14	-0.50	-0.50	2.0	-0.50	2.0	0.10
15	-0.75	-0.75	2.0	-0.75	2.0	0.10

2.3. Estimation of the Parameters

Estimation of the parameters was done by using the Markov Chain Monte Carlo (MCMC) method as implemented in the computer software OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009). A burn-in period of 3,000 iterations was used with a total number of 30,000 iterations for each model. Following priors were used for MCMC estimation of model parameters:

$$\begin{gathered} b_i \sim \text{Normal}(0,1), \qquad i = 1, \dots, n, \\ a_i \sim \text{Normal}(0,1) \text{ and } a_i > 0\,,\; i = 1, \dots, n, \\ c_i \sim \text{Beta}(5,17) \text{ and } 0 < c_i < 0.3, \quad i = 1, \dots, n. \end{gathered} \tag{4}$$

Following priors were used for estimation of ability parameter, depending on the prior assumptions regarding the ability:

$$\theta_j \sim \text{Normal}(0,1), \qquad j = 1, \dots, N,$$

or (5)

$$\theta_j \sim \text{Uniform}(-4,4), \qquad j = 1, \dots, N.$$

The scale of ability is arbitrary in IRT estimation which is denoted as metric identification problem (de Ayala, 2009, p. 41; Baker & Kim, 2004). The metric of the ability requires to be identified to achieve comparable parameter estimates across different calibrations. In this study, the metric of the ability was identified using item centering method (de Ayala, 2009). That is, the mean of item difficulty parameter estimates were fixed to zero for estimation of each model. In addition, the scale of parameters from estimated models was placed on scale of the generating parameters by using mean and sigma equating method (Marco, 1977).

2.4. Item Recovery Analyses

Item recovery analyses were conducted to compare the generating parameters to the parameter estimates from the MCMC analyses with a normal prior and the MCMC analyses with a uniform prior. Accuracy indices and Pearson correlations were calculated for this purpose. The accuracy indices included mean bias, mean absolute error (MAE), mean-square error (MSE), and root-mean-square error (RMSE). The mean bias, MAE, MSE, RMSE and Pearson correlation values were calculated across twenty-five replications for the 15-item and 600 sample size condition, and for the 30-item and 2,000 sample size condition, individually, for each IRT model. As an example, the equations for calculating the accuracy indices and Pearson correlation for the item difficulty parameter (b) are given below:

$$\text{Bias}(\hat{b}) = \frac{\sum_{r=1}^{R}\sum_{i=1}^{n}(\hat{b}_i - \hat{b}_{ir})}{Rxn}, \tag{6}$$

$$\text{MAE}(\hat{b}) = \frac{\sum_{r=1}^{R}\sum_{i=1}^{n}|\hat{b}_i - \hat{b}_{ir}|}{Rxn}, \tag{7}$$

$$\text{MSE}(\hat{b}) = \frac{\sum_{r=1}^{R}\sum_{i=1}^{n}(\hat{b}_i - \hat{b}_{ir})^2}{Rxn}, \tag{8}$$

$$\text{RMSE}(\hat{b}) = \sqrt{\frac{\sum_{r=1}^{R}\sum_{i=1}^{n}(\hat{b}_i - \hat{b}_{ir})^2}{Rxn}}, \tag{9}$$

$$\text{Cor}(\hat{b}, b) = \frac{1}{R}\sum_{r=1}^{R}\text{Cor}(\hat{b}_i, \hat{b}_{ir}), \tag{10}$$

where ($\hat{b}_i$) is the generating item difficulty parameter for item i, ($\hat{b}_{ir}$) is the item difficulty parameter estimate for item i from MCMC analyses with a uniform/normal prior from the rth replication, R is total number of replications which is 25, and n is the total number of items which is either 15 or 30.

3. RESULT / FINDINGS

The accuracy indices and the correlation coefficients are calculated for the item difficulty, item discrimination, item pseudo-guessing, and ability to quantify the item parameter recovery (see Appendix A, Tables A1-A4). Post-hoc comparisons were conducted for transformed MSE values using Tukey's HSD procedure (see Table 2). Square-root or natural logarithm transformation was used for transformation of the MSE values in order to achieve normally distributed residuals. Cohen's d values for the post-hoc comparisons are reported in Table 2. Cohen's d values of 0.2, 0.5, and 0.8 indicate small, medium, and large effects, respectively (Cohen, 1988). Cohen's d values of 0.8 and larger were considered to reveal a substantial difference in mean MSE values between the uniform and the normal priors for a given parameter from a particular model for a given number of the items and the sample size condition.

Results did not indicate a difference in the mean MSE values between the normal and the uniform priors for the item difficulty parameter from the Rasch model, for both the 15-item and 600 sample size and for the 30-item and 2,000 sample size conditions. There was not a constant pattern for differences in the mean MSE values between the uniform and the normal priors for the ability parameter from Rasch model.

For the 15-item and 600 sample size conditiom, there was not a substantial difference in the mean MSE values between the uniform and the normal priors for estimation of the item difficulty and the item discrimination parameters using a 2PL model, when the actual distribution of the latent ability was uniform. When the actual distribution of the latent ability was normal, the uniform prior yielded larger mean MSE value compared to the normal prior for both of the item difficulty and item discrimination parameters. For the 30-item and 2,000 sample size condition, for both of the item difficulty and item discrimination parameters, the normal prior yielded larger mean MSE value when the actual distribution of the latent ability was uniform. Similarly, the uniform prior yielded larger mean MSE value when the actual distribution of the latent ability was normal, for both of the item difficulty and item discrimination parameters. For estimation of the ability parameter using a 2PL model, the normal prior yielded larger mean MSE values compared to the uniform prior for all conditions.

The analyses of the 15-items using a 3PL model for 600 sample size showed that, there was not a substantial difference in the mean MSE values between the uniform and the normal priors for the item difficulty and the item discrimination parameters, when the actual distribution of the latent ability was uniform. For the item pseudo-guessing parameter, the uniform prior yielded larger errors compared to the normal prior, when the actual latent ability distribution was uniform, for the 15-item and 600 sample size condition. Again for the 15-item and 600 sample size condition, the uniform prior yielded larger errors compared to the normal prior, when the actual latent ability distribution was normal, for estimation of the item difficulty, item discrimination and item pseudo-guessing parameters.

For the 30-item and 2,000 sample size condition, the normal prior yielded larger mean MSE values compared to the uniform prior, for estimation of the item difficulty and item pseudo-guessing parameters, when the actual latent ability distribution was uniform. For the item discrimination parameter, on the other hand, there was not a significant difference in the mean MSE values between the normal and uniform priors. Again for the 30-item and 2,000 sample size condition, the uniform prior yielded larger mean MSE values for the item difficulty, item discrimination, and item pseudo-guessing parameters, when the actual distribution of the latent ability was normal. For estimation of the ability parameter in the 3PL model, the normal prior yielded larger mean MSE values compared to the uniform prior, when the actual latent ability distribution was uniform. The effect sizes for the difference between the normal and uniform priors were medium to large (i.e., between 0.5 and 0.8) when the actual distribution was normal.

Table 2. Estimates of Cohen's d Values from Post-hoc Comparisons Using Tukey'd HSD Procedure for Transformed MSE Values

			Rasch		**2PL**			**3PL**			
Condition	**Actual Dist.**	**Prior Dist.**	***b***	***θ***	***b***	***a***	***θ***	***b***	***a***	***c***	***θ***
15-item and 600 sample size	Uniform	Normal – Uniform	0.138 U>N	**1.240 U>N**	0.579 N>U	0.236 U>N	**3.467 N>U**	0.611 N>U	0.780 U>N	**0.915 U>N**	**7.627 N>U**
	Normal	Normal – Uniform	0.026 U>N	0.455 U>N	**1.243 U>N**	**2.819 U>N**	**7.526 N>U**	**2.299 U>N**	**5.090 U>N**	**2.319 U>N**	0.756 N>U
30-item and 2,000 sample size	Uniform	Normal – Uniform	0.245 U>N	**3.809 U>N**	**0.999 N>U**	**1.314 N>U**	**3.197 N>U**	**2.240 N>U**	0.281 U>N	**1.466 N>U**	**6.022 N>U**
	Normal	Normal – Uniform	0.013 U>N	**2.092 N>U**	**1.231 U>N**	**3.914 U>N**	**3.903 N>U**	**2.998 U>N**	**7.894 U>N**	**1.056 U>N**	0.727 U>N

Note. 1) Dist: Distribution, N: Mean parameter estimates for the model with normal prior, U: Mean parameter estimates for the model with uniform prior, *b*: Item difficulty, *a*: Item discrimination, *c*: Item pseudo-guessing, *θ*: Ability 2) Large effect sizes (i.e., larger than .80) are shown in bold.

4. DISCUSSION and CONCLUSION

The primary purpose of using IRT models is to locate students on a continuous scale by estimating their ability (Baker, 2001). Thus, correct estimation of the ability parameters in an IRT model is critical for accountability. The purpose of this study was to investigate if the prior distribution assumption for ability has an effect on estimation of the ability parameters, especially when the true ability distribution is uniform. For this purpose, a simulation study was conducted to compare a uniform and a normal prior distribution assumption for Bayesian estimation of the item and ability parameters in Rasch, 2PL and 3PL models. The simulation conditions included a short test with a small sample size, and a long test with a large sample size. Ability distributions were generated to follow either a normal or a uniform distribution; and the item responses were generated to fit either a Rasch, 2PL or a 3PL model. Twent-five data sets of item responses were generated for each combination of the simulation conditions. Each data set was analyzed using both a uniform and a normal prior, and the ability and item parameter estimates from both models were compared for their accuracy.

Uniform and normal priors for ability yielded similar item parameter estimates for the Rasch model for each simulation conditions. The uniform and normal priors either resulted similar item parameter estimates, or the prior that does not match the true distribution resulted in better estimates of the ability parameter. That is, a uniform distribution prior yielded more accurate estimates of the ability parameter when the true distribution of ability was normal; and the normal prior resulted in more accurate ability parameter estimates when the true distribution of ability was uniform.

For the 2PL model, the normal and the uniform distribution priors for ability either resulted in similar item difficulty and item discrimination parameter estimates, or the prior distribution that matches the true distribution of ability resulted in more accurate estimates of similar item difficulty and item discrimination parameters. For estimation of the ability parameters, the uniform distribution prior yielded more accurate estimates for each of the simulation condition independent of the true distribution of ability.

The uniform and normal distribution priors for ability either resulted in similar item difficulty parameter estimates, or the prior that matches the true distribution of ability yielded more accurate item difficulty parameter estimates. Uniform and normal distribution priors resulted in similar item discrimination parameter estimates when the true distribution was uniform. Normal distribution prior yielded more accurate estimates of the item discrimination parameter when the true distribution of ability was normal. Similarly, the normal distribution prior yielded more accurate estimates of the item pseudo-guessing parameter for each of the simulation conditions except for the 30-item and 2,000 sample size condition, when the true distribution of ability was uniform. For this condition, the uniform distribution prior resulted in more accurate estimates of the item pseudo-guessing parameters. The uniform and normal distribution priors for ability yielded similar estimates of ability when the true distribution of ability was normal. When the true distribution of ability was uniform, on the other hand, the uniform distribution prior yielded more accurate estimates of the ability parameter.

In summary, the results of this study suggest using a uniform distribution prior to achieve more accurate estimates of the ability parameter in the 2PL and 3PL models when the true distribution of ability is not known. The results contribute to the IRT literature as they suggest that using a uniform prior for ability may be more useful as opposed to the convention of using a normal prior for estimation of ability. The results did not indicate a guiding pattern for estimation of the ability parameter in the Rasch model. However, the results of this study are limited with the simulation conditions used in the study. A future study may include more alternatives for the test length and the sample size conditions. In addition, this study only investigated the effect of the prior distribution for ability on estimation of the parameters in IRT models. A future study may explore potential effects of the prior distributions for the item parameters on parameter estimation in IRT models.

Acknowledgements

An earlier version of this paper was presented at the International Conference on Research in Education and Science (ICRES), Kusadasi, Aydin, Turkey, May 18-21, 2017.

ORCID

Tuğba Karadavut https://orcid.org/0000-0002-8738-7177

5. REFERENCES

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland. Retrieved from http://files.eric.ed.gov/fulltext/ED458219.pdf

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 397-479). Reading, MA: Addison-Wesley. Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, D. L. (1959). A replication of Lord's study on skewness and kurtosis of observed test-score distributions. *Educational and Psychological Measurement, 19*, 81-87.

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

de Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement, 23*, 3-19.

Embretson, S. E. (1996). The new rules of measurement. Psychological Assessment, 8, 341.

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

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357-381.

Hambleton, R. K., & Cook, L. L. (1983). *The robustness ofitem response models and the effects of test length and sample size on the precision of ability estimates*. In D. Weiss (Ed.), *New horizons in testing* (pp. 31–49). NewYork: Academic Press.

Jackman, S. (2000). Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. *American Journal of Political Science, 44*, 375-404.

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*, 146–162.

Lord, F. M. (1955). A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement, 15*, 383-389.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores* (with contributions by A. Birnbaum). Reading, MA: Addison-Wesley.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in medicine, 28*, 3049-3082.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

Micerri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved January 10, 2017, from https://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielson and Lydiche (for Danmarks Paedagogiske Institut).

Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement, 26*, 192-207.

Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education, 21*, 65-88.

Sen, S., Cohen, A. S., & Kim, S.-H. (2016). The impact of non-normality on extraction of spurious latent classes in mixture IRT models. *Applied Psychological Measurement, 40*, 98-113.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299-311.

Stewart, J. (2012) Does IRT provide more sensitive measures of latent traits in statistical tests? An empirical examination. *Shiken Research Bulletin, 16*, 15-22.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). *Assessing the fit of item response theory models*. In C. R. Rao & S. Sinharay (Eds.), Psychometrics: Vol. 26. Handbook of statistics (pp. 683–718). Amsterdam: Elsevier.

6. APPENDIX

Table A1. Accuracy Indices and Correlations for the 15-item and 600 Sample Size Condition when the Actual Latent Ability Distribution is Uniform

	MRM				2PL					
	Item difficulty		Ability		Item difficulty		Item discrimination		Ability	
Prior	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal
Bias	0.000	0.000	1.037	0.977	0.000	0.000	0.036	0.030	0.900	0.932
MAE	0.083	0.086	1.078	1.011	0.073	0.080	0.124	0.114	0.916	0.949
MSE	0.011	0.012	1.506	1.405	0.009	0.011	0.025	0.022	1.064	1.145
RMSE	0.107	0.109	1.227	1.185	0.092	0.107	0.159	0.149	1.031	1.070
Cor.	0.995	0.995	0.931	0.930	0.996	0.995	0.945	0.961	0.957	0.953

Table A1 Continues. Accuracy Indices and Correlations for the 15-item and 600 Sample Size Condition when the Actual Latent Ability Distribution is Uniform

	3PL							
	Item difficulty		Item discrimination		Item pseudo-guessing		Ability	
Prior	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal
Bias	0.000	0.000	-0.123	-0.151	0.001	-0.018	0.922	1.039
MAE	0.124	0.136	0.234	0.205	0.035	0.030	0.978	1.059
MSE	0.026	0.033	0.078	0.061	0.002	0.001	1.295	1.477
RMSE	0.162	0.183	0.280	0.246	0.043	0.038	1.138	1.215
Cor.	0.989	0.986	0.869	0.953	0.735	0.868	0.933	0.930

Table A2. Accuracy Indices and Correlations for the 15-item and 600 Sample Size Condition when the Actual Latent Ability Distribution is Normal

	MRM				2PL					
	Item difficulty		Ability		Item difficulty		Item discrimination		Ability	
Prior	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal
Bias	0.000	0.000	0.848	0.901	0.000	0.000	-0.172	-0.047	0.862	0.978
MAE	0.078	0.078	0.918	0.921	0.147	0.104	0.257	0.154	0.876	0.983
MSE	0.010	0.010	1.144	1.110	0.033	0.021	0.099	0.038	0.969	1.145
RMSE	0.099	0.099	1.069	1.054	0.182	0.145	0.314	0.195	0.985	1.070
Cor.	0.996	0.996	0.833	0.835	0.986	0.991	0.790	0.900	0.900	0.908

Table A2 Continues. Accuracy Indices and Correlations for the 15-item and 600 Sample Size Condition when the Actual Latent Ability Distribution is Normal

	3PL							
	Item difficulty		Item discrimination		Item pseudo-guessing		Ability	
Prior	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal
Bias	0.000	0.000	-0.114	-0.150	0.020	0.001	0.967	1.062
MAE	0.230	0.136	0.447	0.210	0.049	0.040	1.007	1.070
MSE	0.075	0.035	0.288	0.061	0.004	0.002	1.389	1.407
RMSE	0.274	0.187	0.537	0.247	0.061	0.048	1.179	1.186
Cor.	0.968	0.985	0.226	0.913	0.432	0.656	0.868	0.875

Table A3. Accuracy Indices and Correlations for the 30-item and 2,000 Sample Size Condition when the Actual Latent Ability Distribution is Uniform

	MRM				2PL					
	Item difficulty		Ability		Item difficulty		Item discrimination		Ability	
Prior	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal
Bias	0.000	0.000	1.136	1.000	0.000	0.000	-0.002	0.037	1.003	1.029
MAE	0.048	0.050	1.145	1.005	0.046	0.057	0.063	0.079	1.006	1.034
MSE	0.004	0.004	1.548	1.249	0.004	0.006	0.007	0.011	1.156	1.235
RMSE	0.059	0.062	1.244	1.117	0.060	0.075	0.081	0.104	1.075	1.112
Cor.	0.998	0.998	0.958	0.960	0.998	0.998	0.982	0.985	0.975	0.971

Table A3 Continues: Accuracy Indices and Correlations for the 30-item and 2,000 Sample Size Condition when the Actual Latent Ability Distribution is Uniform

	3PL							
	Item difficulty		Item discrimination		Item pseudo-guessing		Ability	
Prior	Uniform	Normal	Uniform	Normal	Uniform	Normal	Uniform	Normal
Bias	0.000	0.000	-0.172	-0.036	-0.002	-0.019	0.987	1.043
MAE	0.076	0.113	0.207	0.188	0.019	0.025	0.999	1.049
MSE	0.010	0.023	0.059	0.052	0.001	0.001	1.222	1.357
RMSE	0.099	0.153	0.242	0.229	0.025	0.033	1.106	1.165
Cor.	0.996	0.990	0.928	0.973	0.919	0.910	0.958	0.953

Table A4. Accuracy Indices and Correlations for the 30-item and 2,000 Sample Size Condition when the Actual Latent Ability Distribution is Normal

	MRM				2PL					
	Item difficulty		Ability		Item difficulty		Item discrimination		Ability	
Prior	**Uniform**	**Normal**	**Uniform**	**Normal**	**Uniform**	**Normal**	**Uniform**	**Normal**	**Uniform**	**Normal**
Bias	0	0	0.838	0.945	0	0	-0.117	-0.014	0.945	1.003
MAE	0.047	0.047	0.863	0.948	0.090	0.061	0.158	0.076	0.946	1.003
MSE	0.003	0.003	0.935	1.071	0.013	0.009	0.037	0.009	1.023	1.115
RMSE	0.059	0.059	0.967	1.035	0.115	0.093	0.191	0.095	1.011	1.056
Cor.	0.999	0.999	0.904	0.906	0.994	0.996	0.956	0.976	0.935	0.944

Table A4 Continues. Accuracy Indices and Correlations for the 30-item and 2,000 Sample Size Condition when the Actual Latent Ability Distribution is Normal

	3PL							
	Item difficulty		Item discrimination		Item pseudo-guessing		Ability	
Prior	**Uniform**	**Normal**	**Uniform**	**Normal**	**Uniform**	**Normal**	**Uniform**	**Normal**
Bias	0.000	0.000	0.032	-0.032	0.025	0.005	0.932	0.968
MAE	0.182	0.088	0.470	0.145	0.036	0.029	0.941	0.969
MSE	0.047	0.016	0.364	0.034	0.002	0.001	1.101	1.087
RMSE	0.216	0.126	0.604	0.185	0.043	0.038	1.049	1.043
Cor.	0.980	0.993	-0.058	0.922	0.829	0.806	0.910	0.921

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 580–591
https://dx.doi.org/10.21449/ijate.559678

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

Assessing time knowledge in children aged 10 to 11 years

Nicola Brace [1], **Clare Doran** [2], **Janet Pembery** [3], **Emma Fitzpatrick** [1],
Rosalind Herman [1,*]

[1]City, University of London, School of Health Sciences, Division of Language and Communication Science UK
[2]Bedfordshire Community Health Service, UK
[3]Central and North West London NHS Foundation Trust, UK

ARTICLE HISTORY

Received: 01 May 2019

Revised: 13 Septembe 2019

Accepted: 26 October 2019

KEYWORDS

Time knowledge,
Assessment,
Clock reading,
Time orientation,
Estimating time intervals

Abstract: The acquisition of time knowledge involves learning how to read clocks, estimate time, read dates and learn about temporal sequences. Evidence suggests that many of these competencies are acquired by 10 years of age although not all children may follow this developmental path. The main purpose of this study was to collect normative data for a screening tool that assesses time knowledge. These data identify the prevalence and pattern of difficulties with time knowledge among a UK sample of Year 6 pupils (aged 10 to 11 years). The Time Screening Assessment tool (Doran, Dutt & Pembery, 2015), designed to assess time knowledge, was administered individually to a sample of 79 children. Findings revealed a median overall score of 32 out of a maximum score of 36. 25% of children performed at or close to ceiling, however seven children scored more than –1.5 standard deviations below the mean. The value of these findings to practitioners working with children in schools is discussed.

1. INTRODUCTION

In everyday life, keeping track of time allows us to organise activities and coordinate these with others, and is a skill that is acquired during childhood. Burny, Valcke and Desoete (2009) suggested that what develops is a range of time-related competencies including the accurate reading of clocks and calendars, and the ability to use mental timelines to measure and estimate time intervals. Furthermore, Burny et al. highlighted that the specific skill of reading clocks draws on a number of sub-competencies including language skills, memory, numeracy and spatial abilities. They explained that as well as being able to count and have a basic understanding of fractions, children need to learn to express correctly the relationship between the hour and the minute. For the relative expression 'ten past eleven' the minute is mentioned before the hour; however, for the absolute expression 'eleven ten' the hour comes first.

Clock reading has been explored in a number of studies. Siegler and McGilly (1989) concluded from North American studies conducted in the 1970s and 1980s that children develop the ability to tell the time from analogue clocks in a particular sequence. By 6 years of age many can tell

CONTACT: Rosalind Herman ✉ R.C.Herman@city.ac.uk Division of Language and Communication Science, School of Health Sciences, City, University of London, Northampton Square, EC1V 0HB

ISSN-e: 2148-7456 /© IJATE 2019

whole-hour times, by 7 or 8 years 5-minute times, and between the ages of 8 and 10 years many could tell 1-minute times. However, reading the time from digital clocks does not follow the same pattern. A study by Friedman and Laycock (1989) involved participants from five age groups, with 32 North American children in each (mean ages: 6;6, 7;7, 8;6, 9;7 and 10;6). Most in the youngest group could tell whole-hour, half-hour and 1-minute times when reading a digital clock, and performance was near perfect in the second age group. However, their performance when reading analogue clocks varied according to the time being displayed. Whole-hour readings were accurate in the youngest age group and half-hour times in the second age group, but more complex 1-minute times (such as 2:43) remained difficult for at least some children in the oldest age group. This early research suggests that analogue clock reading is a complex skill acquired over a period of time, but in the most part it is achieved by 10 years of age.

Several studies have shown that numerical skills may affect the acquisition of clock reading skills. Andersson (2008) compared clock reading in 182 children attending school in Sweden with a mean age of 125 months. Those with mathematics difficulties were found to have substantial problems with reading both analogue and digital clocks. Burny, Valcke and Desoete (2012) sampled 725 children from eight Belgian primary schools and identified 154 children with mathematics difficulties who performed worse than the others on clock reading tasks; furthermore, telling the time accurately to 1-minute and 5-minutes was difficult with both analogue and digital clocks. Analysis of errors suggested both miscounting and misinterpreting, with the latter most likely due to a combination of difficulties, including poorer counting strategies and absent memory representations. For example, reporting 10:04 rather than 10:20 suggests that the child was not counting in fives, reflecting a lack of knowledge that the ‘4’ on the analogue clock means ‘20’.

Clock reading is just one aspect of time knowledge. Other research has focused on the acquisition of knowledge about temporal sequences. In a US study, Friedman (1991) looked at children’s ability to date events on a time scale. The children were aged 4, 6 and 8 years with 14 children in each age group. They were asked about two events they had experienced, one staged seven weeks and the other one week prior to testing. The youngest group could accurately decide which of the two events was more recent, and therefore had a sense of different times in the past, but it was only the 6 and the 8-year old children who could estimate when the older event occurred and who showed awareness of day, month and season. In a further study, Friedman (1992) compared the same three age groups, asking children to recall events from specific points during the previous year and, whilst performance improved with age, only 56% of those in the oldest group could position these multiple events into their correct temporal order. These findings suggest that acquisition of time-awareness continues beyond the age of 8 years. Based on these and other findings, Friedman (2005) proposed that children first learn the order of the days of the week and months of the year, using a list-based representation. As they grow older they begin to form representations of longer time scales such that by 10 years of age they have a sense of the annual cycle, for example they can judge how long it is until next summer, and become aware of temporal distances between the days of the week or months of the year, for example that April and October are quite far apart.

Another component of time knowledge concerns the ability to judge how long something takes. A study by Quartier, Zimmermann and Nashat (2010) compared Swiss French-speaking children aged between 6 and 13 years, 22 with attention-deficit/hyperactivity disorder (ADHD) and 22 controls. They found that children with ADHD who were younger than 10 years of age had more difficulty than controls with conventional time concepts such as dates, durations and order of events. Although those older than 10 years of age showed conventional time knowledge, they differed significantly from the control group in terms of their ability to organise time, for example to plan forward and meet deadlines. Children with autism spectrum condition (ASC) have also displayed difficulties with time-based judgements (e.g. Williams, Boucher, Lind & Jarrold, 2013).

A number of researchers have developed questionnaires to measure the different components of time knowledge. Quartier et al. (2010) used the Time Concept Questionnaire (Quartier, 2008) consisting of questions relating to time orientation, conventional time sequences, objective durations, subjective durations and anticipation. Labrell, Mikaeloff, Perdry and Dellatolas (2016) developed their own Time Knowledge Questionnaire which included four subtests: time orientation (e.g. 'What day is it today?), sequences in relation to months and seasons, time units (e.g. 'Is a minute shorter or longer than a second?') and telling the time on a clock. There were three other subtests designed to measure understanding of the lifespan, the child's own birthday and time estimation. The latter was assessed through a question about the duration of the interview. They administered this to 105 French children from state schools, ranging in ages from 6 to 11 years, and found that although time knowledge increased with age, different subtests revealed different patterns. For example, time orientation was at ceiling from 7 years of age, whereas time estimation continued to improve between 9 and 10 years of age. Furthermore, when controlling for age they found significant correlations between some subtests but not all and suggested that what they were measuring might not be unidimensional.

Whereas Labrell et al. (2016) were concerned with the development of time knowledge between the ages of 6 to 11 years, Dutt and Doran (2013) reported data using a similar questionnaire from 20 young people, aged 13 to 17 years, who had been referred for assessment or therapy to a Youth Offending Team. They found nine young people had difficulties associated with estimating and telling the time, with calendar time (i.e. naming the months in the correct order and interpreting a short date), and with understanding the word 'fortnight'. Importantly, these findings are in contrast to the research evidence presented so far which suggested that time knowledge competencies are acquired fully by around 10 years.

The questionnaire used in the Dutt and Doran study has since been published along with a resource pack (Doran, Dutt and Pembery, 2015). It is called the Time Screening Assessment and was developed because of the authors' experiences as Speech and Language Therapists. As therapists they found some young people to have a poor sense of time, they were either missing appointments or were late, and had difficulties with temporal sequences and clock reading. Colleagues working in secondary schools had highlighted similar difficulties among some pupils aged older than 10 years. The Time Screening Assessment was developed as a tool to allow the identification of children who are not acquiring time knowledge according to the usual developmental trajectory and assesses knowledge that is taught in schools in England before the age of 10 years. According to the UK National Curriculum Statutory Guidance (2013) it is a statutory requirement for pupils to have been taught by nine years of age clock reading skills (including from analogue clocks), temporal sequences (identifying chronological order using language, and recognising and using language related to dates) and also estimating time and comparing the durations of events.

The aim of the present study was to explore the incidence of poor time knowledge in a non-clinical group of children aged 10 to 11 years and thereby provide normative data for the Time Screening Assessment. Based on the findings of previous research, suggesting that time-related competencies are achieved by 10 years of age, the majority of scores were predicted to be at ceiling, and based on Dutt and Doran (2013) it was also predicted that some participants might not score highly on this assessment tool. No predictions were made with regards to the possible effects of gender, type of school or age, however these were explored when analysing overall performance. Based on the findings of Labrell et al. (2016), it was also predicted that there would be significant correlations between different sections of the questionnaire.

2. METHOD

2.1. Design of the study

This study used a questionnaire designed to measure time knowledge.

2.2. Participants

This study received ethical approval from City, University of London. In order to obtain a sample of 10 to 11-year olds representative of those attending state schools in South East England, five different schools were approached that were located in Buckinghamshire and Greater London. Participants were recruited from Year 6 of five government-funded primary schools, two of which were located in a village (with 30 and 88 pupils), two in a town (with 45 and 60 pupils) and one in a large city (with 90 pupils). At village and town schools the proportion of pupils for whom the school received a pupil premium (additional funding for disadvantaged children) was below average, as was the proportion of pupils who had special educational needs. The city school had an above average proportion of pupils receiving the pupil premium, with special educational needs, and with English as an additional language.

All year 6 pupils at each school took home information about the study. In compliance with ethics approval, parents were invited to return a signed consent form to the Year 6 teacher. Parents were also asked to provide optional information, including their child's date of birth (in order to accurately score one question), whether their child had received a diagnosis of ADHD/ASC, and whether they had received any speech and language therapy.

Of the 81 children for whom a consent form was returned 37 were girls and 44 were boys. They were aged between 123 and 136 months (mean age = 129.74 months, SD = 3.19). The final sample included in data analysis comprised 32 pupils from village schools, 34 from town schools and 13 from a city school. The criteria for including data in the analysis were that participants did not have a diagnosis of ADHD or ASC as both conditions have been linked to difficulties with time-based judgements. Two children (boys) did not fit the inclusion criteria and their responses were excluded from data analysis.

2.3. Materials

The Time Screening Assessment (Doran et al., 2015) has five sections, with multiple questions in each: Calendar time; Clock time; Time vocabulary; Organisation of time; and Estimation of time. In total, there are 25 questions, with the majority requiring a response that is either correct or incorrect (e.g. What does 'fortnight' mean? In which month is Christmas? What is the time shown here?). Four questions ask respondents to indicate a strategy (e.g. How do you know when it is time to get up in the morning?) and four questions ask for an estimate of time duration (e.g. Approximately how many minutes does a song on the radio and a school lesson last?).

Following the advice of the authors of this measure, and based on their experience of using the tool, three questions were amended to suit the age group and diversity of cultures of the participants: 'Explain exactly what each number means in this date' was amended to 'What date is this?'; 'How long do you think this assessment has taken?' was amended to 'How long has it taken to answer these questions'; and 'Which season is usually hot?' was amended to 'Which season is usually hot here in England?'.

Three images were printed on A4 paper for the purposes of asking three of the questions: a date in a short format (03/06/12); a digital clock showing 7:20; and an analogue clock showing 11:05. The validity and reliability of the assessment tool has not to date been evaluated, although it does have face validity as it assesses time knowledge taught according to the UK National Curriculum Statutory Guidance. Investigation of scorer reliability was carried out as part of this study.

2.4. Procedure

All participants were assessed individually in school by the primary author. Children were invited to sit at a quiet desk outside of their classroom. An analogue clock was placed on the desk. Each question from the tool was read out and their response was recorded in writing, either verbatim or précised. Positive encouragement was provided throughout, regardless of whether responses were correct, and assessments lasted between 10 and 15 minutes. At the end of the session children were asked if they found any of the questions difficult. Responses to this final question were not scored or included in the analysis, but reassurance was provided if any concerns were raised.

2.5. Scoring

Scoring followed the guidance provided by Doran et al. (2015). Four questions in Calendar time were coded 0 (incorrect) or 1 (correct). In terms of providing today's date, participants only had to specify the correct day and month, and for the date of their birthday they could be prompted to provide the year. When asked to name the seasons 1 point was awarded for naming all four seasons and a further point for naming them in the correct order. When asked to explain a short date 1 point was awarded for correct naming of each of the day, month and year. When asked to name the months of the year in order 3 points were awarded if all 12 months were provided in the correct order, 2 points if one or two errors of omission or order, 1 point if three errors and 0 points if four or more errors.

Three questions in Clock time were coded 0, 1 or 2 points for each clock shown (digital and analogue), with 2 points being were awarded if the time was correctly described using both relative and absolute expressions, 1 point if one of these expressions was used and 0 if the time was not correctly identified. No points were awarded for answers such as '50 past 7' or '35 past 11'. When asked what each clock would be in half an hour, 2 points were awarded if the correct time was provided for both clocks and 1 point if correct for one clock. For the final question 'What is the time now?', responses were given 1 point if correct to within two minutes. The four questions in Time vocabulary were coded as either 0 (incorrect) or 1 (correct), as were the four questions in Organisation of time where 0 was given if the response indicated that the child predominantly relied on another person and 1 if the child used a strategy that did not involve another person.

For Estimation of time, three questions were coded as 0, 1 or 2 points. Each involved asking children to estimate how long two activities lasted (minutes, hours or weeks/months). 1 point was awarded if children estimated the length of a song as 2-5 minutes, 1 point for correct length of a lesson, 1 point for saying the length of a film was between 1¼-3 hours, 1 point for saying the school day was 6-8 hours, 1 point for estimating the length of a term as 12-14 weeks (or half term 6-8 weeks) and 1 point for saying the length of the school summer holiday break was 5-7 weeks. The final three questions in this section were coded as either 0 or 1 point. Reasonable answers to name something that takes an hour to do were scored as 1 point, for example football practice or English homework. Responses when asked to estimate how many months or weeks until their next birthday were given 1 point if correct to within a month. Responses when asked how long the assessment had taken were awarded 1 point if correct to within five minutes.

2.6. Reliability

A sample of 30 assessments was independently scored by a second person who was briefed on the scoring system outlined above. An intra-class correlation coefficient (ICC) was used to assess reliability of the overall score given by each scorer. The reliability coefficient was calculated as .981, with 95% CI (.950, .992) indicating a high level of agreement between the two scorers.

3. RESULTS

3.1. Overall performance on the Time Screening Assessment

The scores for each question for 79 child participants were analysed and an overall score, out of a maximum of 36, was calculated for each participant. The distribution of overall scores, shown in Figure 1, is negatively skewed.

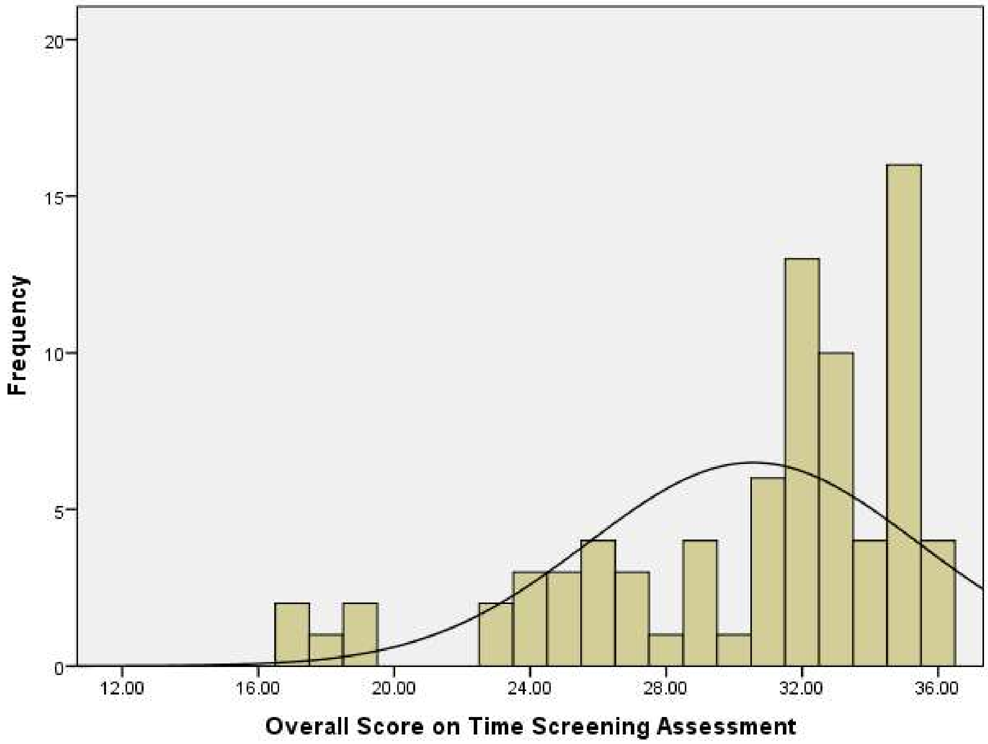


Figure 1. Distribution of overall scores on the Time Screening Assessment

25.4% of the sample scored full (36/36) or almost full marks (35/36). A further 43.2% scored between 30/36 and 34/36. However, 31.6% scored below 30/36. Transformation to z scores revealed that 16.5% were more than –1 standard deviation from the mean, with an overall score lower than 26/36. Seven scores were in excess of –1.5 standard deviations below the mean, with five scores in excess of –2 standard deviations.

Table 1. Descriptive statistics for overall scores on the Time Screening Assessment, broken down by gender

Gender	*N*	Mean	Median	SD	Minimum	Maximum
Boy	42	30.52	32	5.19	17	36
Girl	37	30.59	32	4.50	19	36
Overall	79	30.56	32	4.85	17	36

Table 1 suggests that as a group there was a range in the scores achieved although the majority were towards the top end. The performance of boys overall appears similar to that of the group of girls, and a Mann-Whitney U test confirmed that there was no significant difference (U = 775.50, N_1 = 42, N_2 = 37, p = .988, two-tailed).

Table 2. Descriptive statistics for overall scores on the Time Screening Assessment, broken down by type of school

Type of school	*N*	Mean	Median	SD	Minimum	Maximum
Village	32	32.28	33	3.63	18	36
Town	34	30.44	32	5.37	17	36
City	13	26.61	25	3.85	19	32

Table 2 above suggests that overall the scores of the children attending the school located in a large city were lower than those of the children attending the schools located in the village or town. A Mann-Whitney U test was used to perform three pairwise comparisons (Bonferroni-corrected *p*-value = .017). No statistically significant difference emerged when comparing village and town schools ($U = 453.00$, $N_1 = 32$, $N_2 = 34$, $p = .237$, two-tailed), however the scores for the city school were significantly lower than those from the town schools ($U = 100.00$, $N_1 = 34$, $N_2 = 13$, $p < .005$, two-tailed) and those from the village schools ($U = 42.50$, $N_1 = 32$, $N_2 = 13$, $p < .001$, two-tailed). It is worth nothing, however, that of the seven lowest scoring participants, one came from one of the village schools, four from the town schools (two from each) and two from the city school.

As there was a range in age from 10 years and 3 months to 11 years and 4 months, Spearman's *r* was calculated to explore any relationship between age and overall assessment score. There was a significant positive correlation between age and the overall score ($r_s = .229$, $N = 79$, $p < .05$, two-tailed), however the strength of the correlation is weak and only 5.24% of the proportion of the variation in the overall scores is explained by age.

3.2. Performance in each section of the Time Screening Assessment

In addition to calculating an overall score, a score for each section of the Time Screening Assessment was calculated for each child.

Table 3. Descriptive statistics for each section of the Time Screening Assessment

Section (maximum possible score)	Mean	Median	SD	Minimum	Maximum
Calendar time (12)	10.56	11	1.93	3	12
Clock time (7)	5.84	7	1.85	1	7
Time vocabulary (4)	3.06	3	0.88	1	4
Organisation of time (4)	3.18	3	0.87	1	4
Estimation of time (9)	7.92	8	1.22	4	9

Table 3 shows the median score to be close to or at the maximum possible, suggesting that the majority of children provided accurate responses to each section. Of the seven lowest-scoring participants, all scored below the median in Estimation of time and Calendar time and, except for one participant, Clock time scores were low. In contrast, all but two of the seven achieved the median in the section Organisation of time.

In relation to Calendar time, nearly half the sample achieved the maximum score. While all children knew their own birthday, 10% responded incorrectly when asked what today's date was and about 35% could not identify the day/month/year when shown a short date. Many misidentified the month '06' as July and some did not recognize '12' as being '2012'. Just over a quarter did not name all the months of the year in the correct order and just over a quarter did not name all four seasons in the correct order.

In terms of Clock time, 63% achieved the maximum score, and a further 10% scored 6/7. Approximately one quarter scored 4/7 or less. Whereas all children were able to read correctly the time displayed by an image of a digital clock, using the relative expression 'twenty past seven' and/or the absolute expression 'seven twenty', 11 participants (nearly 14%) did not read the time when shown an image of an analogue clock using either type of expression. Furthermore, whereas only 10 children did not say accurately what the time would be in half an hour when looking at the digital clock, 25 children did not provide a correct response when shown the image of the analogue clock. Finally, 10 children did not provide an accurate response when asked the time.

For the Time vocabulary questions, almost 40% achieved the maximum score. Almost half could not define the word 'fortnight' correctly and just over a third did not know the meaning of the word 'century'. When asked about Organization of time, 43% achieved the maximum score. The remainder responded that they were reliant on another person in relation to one of four scenarios: knowing when it is time to get up in the morning; knowing when it is time to leave for school; how they remembered an important date and an important time.

Finally, 40% achieved the maximum score with Estimates of time durations. Over 90% were able to correctly answer how many minutes a song and a school lesson lasts, and how many hours a film and a school day lasts, and all but one could name something that takes about an hour to do. However, 28% were unable to correctly estimate how many weeks a school term lasts and 15% did not know how many weeks the school summer holiday lasts. Furthermore, 16.5% children could not correctly estimate how long until their next birthday, and 21.5% did not estimate correctly how long the assessment session had lasted.

To explore whether there were significant relationships between the section scores, a series of Spearman's r were calculated. There were significant relationships between most of the section scores, with the exception of Organisation of time. The results are presented in Table 4.

Table 4. Correlations (r_s) between Time Screening Assessment section scores

N = 79	Clock Time	Time vocabulary	Estimation of time	Organisation of time
Calendar time	.457*	.475*	.445*	.104
Clock time		.495*	.342*	.254
Time vocabulary			.452*	.091
Estimation of time				–.027

* $p < .005$ (the Bonferroni-corrected p-value for performing ten correlations)

4. DISCUSSION and CONCLUSION

The primary aim of this study was to use the Time Screening Assessment (Doran et al., 2015) to explore difficulties with time knowledge in children aged 10 to 11. Among 79 Year 6 pupils approximately 25% of children performed at or close to ceiling and a further 43% achieved at least 30/36, with 32% scoring less that 30/36. The large proportion of children achieving high scores on this assessment supports previous research which suggests that by 10 years of age children have acquired the skills of reading clocks (e.g. Freidman & Laycock, 1989) and knowledge about temporal sequences, such as the months of the year (e.g. Friedman, 1992). However, the distribution of scores showed a long tail of scores lower than 30/36 and transformation to z scores indicated that the performance of 16.5% of the sample was in excess of –1 standard deviation from the mean, including seven cases that were more than –1.5 standard deviations away from the mean.

This finding supports the observation of Dutt and Doran (2013) that some young people have poor time knowledge. However, they identified a larger proportion with difficulties, namely nine out of 20 young people aged 13-17, and the disparity in prevalence is likely to reflect differences between the samples. The sample in the Dutt and Doran study comprised young people referred to the Youth Offending Team, and there is evidence showing that a high proportion of youth offenders have language and communication difficulties (e.g. Bryan, Freer & Furlong, 2007).

Poor numerical skills have been found to hamper the acquisition of clock reading skills (e.g. Burny et al., 2012), and when examining the data from the lowest scoring participants, six of the seven cases showed a particularly low score in the section assessing clock reading. However, clock reading was not the only competency impaired among these participants as scores were also low in the sections assessing temporal sequences and time estimation. The finding that performance was poor in sections other than clock reading is consistent with that reported by Dutt and Doran (2013), who also observed difficulties with the order of the months and estimating time in addition to clock reading problems.

There was no significant effect of gender, and only a weak significant correlation was observed between age and overall score achieved on the Time Screening Assessment, with age accounting for around 5% of the variation in the overall scores. This small effect of age may reflect that certain aspects of time knowledge continue to develop beyond 10 years of age. Labrell et al. (2016) looked at the development in time knowledge from age 6 to 10 years and found that judging interview duration continued to improve after age 9. In the present study, 21.5% did not accurately estimate the length of the assessment.

Statistically significant correlations were observed between most, but not all, of the sections of the Time Screening Assessment, in line with the findings of Labrell et al. (2016). That clock reading abilities might correlate with other aspects of time knowledge is consistent with the point made by Burny et al. (2009), namely that clock reading draws on a number of sub-competencies, including memory, numerical, spatial and language skills. The overall scores from the section Organisation of time, designed to assess reliance on others for time organisation, did not correlate with those from the other sections, and this might reflect the fact that children can acquire good time knowledge but nevertheless continue to rely on another to be on time. Alternatively, children may develop strategies to be on time, for example using their phone as an alarm and for reminders, without acquiring a solid knowledge base concerning dates and temporal sequences, or the ability to estimate the duration of events accurately.

When exploring the pattern of errors that children were making, it is worth highlighting that approximately 14% of the children sampled in the present study were not able to accurately read the time shown on the image of an analogue clock, and approximately 13% could not tell the time from a real analogue clock, whereas all were able to read the time displayed on an image of a digital clock. This pattern is consistent with the findings from Friedman and Laycock (1989) who found that although performance reading a digital clock was near perfect before 10 years of age, reading an analogue clock depended on the time being displayed, with whole-hour and half-hour times being easier, and that even the oldest age group of 10 to 11-year olds had difficulty with more complex time such as 2:43. Several studies have established that numerical skills are implicated in accurate clock reading from both analogue and digital clocks (Andersson, 2008; Burny et al., 2012), and the poor performance observed here may in part be attributed to weak numerical skills. As these were not assessed in the present study it was not possible to explore their contribution further.

It is also worth noting that although performance in identifying the day of the week was close to ceiling (when asked about the day after tomorrow and what day it was two days ago), approximately a quarter of the children sampled in the present study were not able to identify all the months of the year in the correct order and about a quarter could not name the four seasons in the correct order. This finding contrasts with that of Friedman (1992) who found that children aged 8 to 9 years showed awareness of the months and seasons and could order the seasons. However, Friedman (2005) suggested that acquisition of time-awareness continues beyond the age of 9 years when children begin to form representations of longer time scales. The findings of the present study suggest that in this sample children had good representations for the relatively short time scale of a week, but that for some children representations of longer time scales were still developing. Consistent with this explanation is the pattern of errors observed in relation to estimating time durations. Performance overall was poorer when children were asked about events of longer durations, such as how many weeks a term lasted compared to shorter durations such as how many minutes a song lasted.

There are two issues to consider when evaluating the contribution of the present study. Firstly, the sample comprised children whose parents actively consented to their participation and therefore may not be representative as it has been argued that the opt-in (active) consent process may result in a reduced sample size and an increased possibility of sampling bias, limiting the validity and generalizability of the study results (see Hollmann & McNamara, 1999). In the present study, there was a low response rate for each of the five schools; the lowest was for the city school with only 14% of parents returning consent forms and the highest was from one town school with a return rate of almost 38%. Also, the sample was limited to one region of England and therefore the findings reported here may not generalize to other regions of England or other countries.

Secondly, at the present time, there is limited data concerning the reliability and validity of the Time Screening Assessment. In the present study, there was a high level of agreement between the two people scoring the responses from 30 children in the present study pointing to inter-rater reliability. Furthermore, there were correlations between most sections of the tool indicative of internal consistency. Further research assessing test-retest reliability is necessary. In terms of validity, the Time Screening Assessment has face validity as it assesses time knowledge taught according to the UK National Curriculum Statutory Guidance, and it is accompanied by practical resources such as worksheets to help teachers and other professionals address gaps in knowledge about time. The sections included in the assessment also overlap with those included in the Time Concept Questionnaire (Quartier, 2008) and the Time Knowledge Questionnaire (Labrell et al., 2016), which is indicative of content validity. However, in relation to construct validity, the scores from one section, Organisation of time, did not correlate with those from other sections. As mentioned previously there are a number of explanations for this which warrant further investigation.

In conclusion, while further research is needed to establish the reliability and validity of the Time Screening Assessment, the present study provides normative data which is the first step towards creating a standardized, norm-referenced assessment tool. The present study identified that two-thirds of the 79 pupils in this sample of 10 to 11-year olds had well-developed time related competencies, however seven pupils had not acquired the time knowledge that they would be expected to have at their age. As a result, they may later experience difficulties when talking about time and with organising their activities. Time-related skills are valuable as children get older, for example: when they are more likely to be responsible for getting themselves to school; where time concepts appear across different topics in the curriculum; where good time organisational skills are needed to complete the increasing amounts of homework on time and are associated with performing well in examinations. The Time

Screening Assessment can be used by teachers and other professionals to identify children and young people with poor time knowledge so that they can receive targeted support.

ORCID

Nicola Brace https://orcid.org/0000-0003-2928-7327

Rosalind Herman https://orcid.org/0000-0001-5732-9999

5. REFERENCES

Andersson, U. (2008). Mathematical competencies in children with different types of learning difficulties. *Journal of Educational Psychology, 100,* 48–66. http://dx.doi.org/10.1037/0022-0663.100.1.48.

Bryan, K., Freer, J. & Furlong, C. (2007). Language and communication difficulties in juvenile offenders. *International Journal of Language & Communication Disorders, 42,* 505–520. http://dx.doi.org/10.1080/13682820601053977.

Burny, E., Valcke, M. & Desoete, A. (2009). Towards an agenda for studying learning and instruction focusing on time-related competences in children. *Educational Studies, 35,* 481–492. http://dx.doi.org/10.1080/03055690902879093.

Burny, E., Valcke, M. & Desoete, A. (2012). Clock reading: An underestimated topic in children with mathematics difficulties. *Journal of Learning Disabilities, 45,* 351–360. http://dx.doi.org/10.1177/0022219411407773.

Doran, C., Dutt, S. & Pembery, J. (2015). *Time Matters: A Practical Resource to Develop Time Concepts and Self-Organisational Skills in Older Children and Young People.* London: Speechmark.

Dutt, S. & Doran, C. (2013). The meaning of time. *Bulletin: the official magazine of the Royal College of Speech & Language Therapists,* February, 11.

Friedman, W.J. (1991). The development of children's memory for the time of past events. *Child Development, 62,* 139–155. http://dx.doi.org/10.2307/1130710.

Friedman, W.J. (1992). Children's time memory: The development of a differentiated past. *Cognitive Development, 7,* 171-187. http://dx.doi.org/10.1016/0885-2014(92)90010-O.

Friedman, W.J. (2005). Developmental and cognitive perspectives on humans' sense of the times of past and future events. *Learning and Motivation, 36,* 145–158. http://dx.doi.org/10.1016/j.lmot.2005.02.005.

Friedman, W.J. & Laycock, F. (1989). Children's analog and digital clock knowledge. *Child Development 60,* 357–371. http://dx.doi.org/10.2307/1130982.

Hollmann, C.M. & McNamara, J.R. (1999). Considerations in the use of active and passive parental consent procedures. *The Journal of Psychology, 133,* 141–156. http://dx.doi.org/10.1080/00223989909599729.

Labrell, F., Mikaeloff, Y., Perdry, H. & Dellatolas, G. (2016). Time knowledge acquisition in children aged 6 to 11 years and its relationship with numerical skills. *Journal of Experimental Child Psychology, 143,* 1–13. http://dx.doi.org/10.1016/j.jecp.2015.10.005.

Quartier, V. (2008). Le développement de la temporalité: Théorie et instrument de mesure du temps notionnel chez l'enfant [Temporality development: Theory and instrument to measure notional time in children]. *Approche Neuropsychologique des Apprentissages chez l'Enfant, 100,* 76–85.

Quartier, V., Zimmermann, G. & Nashat, S. (2010). Sense of time in children with Attention-Deficit/Hyperactivity Disorder (ADHD). *Swiss Journal of Psychology, 69,* 7–14.

Siegler, R.S. & McGilly, K. (1989). Strategy choices in children's time-telling. In: I. Levin & D. Zakay (Eds.) *Time and human cognition: A life span perspective* (pp.185–218). Amsterdam: Elsevier Science Publishers.

Williams, D., Boucher, J., Lind, S. & Jarrold, C. (2013). Time-based and event-based prospective memory in autism spectrum disorder: The roles of executive function and theory of mind, and time-estimation. *Journal of Autism and Developmental Disorders, 43,* 1555–1567. http://dx.doi.org/10.1007/s10803-012-1703-9.

International Journal of Assessment Tools in Education

2019, Vol. 6, No. 4, 592–601

https://dx.doi.org/10.21449/ijate.554755

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

Factor structure of the CES-D in an impoverished African American sample

Mitchell Porter [1,*], **Youn-Jeng Choi** [2], **Sara Tomek** [3]

[1]Gardner-Webb University, 110 South Main Street, PO Box 7304, Boiling Springs, NC 28017
[2]The University of Alabama, Carmichael Hall, Box 870231, Tuscaloosa, AL 35487
[3]Baylor University, One Bear Place #97301, Waco, TX 76798

ARTICLE HISTORY

Received: 17 April 2019

Revised: 24 July 2019

Accepted: 13 November 2019

KEYWORDS

Factor analysis,
CES-D,
African American,
Impoverished

Abstract: The Center of Epidemiological Studies Depression scale (CES-D) has been used for decades to identify symptomatology of depression in individuals. Overtime, the factor structure of the scale has been both confirmed and challenged when applied to different population samples. The present study explores the factor structure in population sample consisting of impoverished African American parents (*N*=1,020), and the data were collected from the Mobile Youth Study (MYS). Two-, four-, and higher-order models were used to identify the best fitting model. The results indicate that the most parsimonious model is a two-factor structure.

1. INTRODUCTION

The factor structure of the Center for Epidemiological Studies Depression Scale (CES-D) has been challenged through the decades. While Radloff (1977) originally proposed a four-factor structure, other studies have proposed that a two- or three-factor structure may be sufficient (Miller, Markides, & Black, 1997; Manson, Ackerson, Dick, Baron, & Fleming, 1990). In fact, it has even been proposed that one factor solutions may be more desirable for interpretation than multiple factor solutions (Turvey, 1999), and even as many as five-factors has been proposed (Stroup-Benham, Lawrence, & Trevifio, 1992). The purpose of this paper is to investigate the factor structure and find the most parsimonious model of the CES-D within a specific population sample: impoverished African American parents of at-risk behavior students.

There are a number of rating scales designed to measure depression. Some are meant to be completed by researcher, e.g. the Hamilton Depression Scale (Hamilton, 1960) and the Montgomery–Åsberg Depression Rating Scale (Montgomery & Asberg, 1979), while others are designed to be completed by the patients themselves such as the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and Geriatric Depression Scale (Yesavage

CONTACT: Mitchell Porter ✉ mporter1@gardner-webb.edu The University of Alabama, Carmichael Hall, Box 870231, Tuscaloosa, AL 35487

ISSN-e: 2148-7456 /© IJATE 2019

et. al., 1982). In this paper, the CES-D (Radloff, 1977) was used. The scale is comprised of 20 items, divided into four factors (depressed, somatic, positive, and interpersonal). The original proposed factor structure for the CES-D is as follows: depressed affect (blues, depressed, lonely, cry, sad, fearful, failure); positive affect (good, hopeful, happy, enjoy); somatic and retarded activity (bothered, appetite, effort, sleep, get going, mind, talk); and interpersonal (unfriendly, dislike). The CES-D was used due to the high internal consistency in noninstitutionalized adults (i.e. is psychometrically robust) as well as its history of being used in a wide range of populations (Cosco, Prina, Stubbs, & Wu, 2017) making it a good fit for the purposes of this research. The CES-D has historically proven to be one of the most popular forms of assessming depression among researchers (Shafer, 2006).

This four-factor solution has been confirmed in the literature using different sample populations. Husaini, Neff, Harrington, Hughes, and Stone (1980) validated the factor structure in nine rural communities in Tennessee (90% of the sample population was white). Hertzog, Alstine, Usala, Hultsch, and Dixon (1990) confirmed the factor structure in elderly populations (age range 55 to 78) in community-dwelling adults in Canada. Similarly, Lewinsohn, Seeley, Roberts, and Allen (1997) investigated the efficacy of the CES-D as a screener for depression in community-residing older adults (age range = 50 – 96, *N* = 1,006), and concluded that the internal structure of the four-factor model held.

Gender has also been an area of interest. Knight, Williams, McGee, and Olaman (1997) found that the four-factor model fit the data moderately well in a community sample of 675 women in New Zealand; the majority of the sample was white, with 5% being either Chinese or Polynesian. Sommel, Given, Kalaian, Shultz, and McCorkle (1993) explored gender bias in the measurement properties in the CES-D. The results indicated that the four-factor structure held, but exhibited signs of gender bias. They created a subset of 15 items that eliminated gender bias and still captured nearly all the information of the 20-item CES-D. Items excluded were failure, talk, unfriendly, crying, and dislike.

Research on the factor structure of the CES-D in minority populations has also been extensive. Roberts (1980) confirmed the four factor structure in a sample containing Anglos (*n* = 254), Blacks (*n* = 270), and Mexican Americans (*n* = 181). Clark, Aneshensel, Freriches, and Morgan (1981) analyzed the effects of gender and age in response to the CES-D items in a sample of 1,000 adults (61% White, 20% Hispanic, 12% Black, and 7% other; their results showed that while the interitem correlations were significantly higher for women than in men, the four-factor solution was confirmed. Williams et. al (2007) examined and confirmed the factor structure in a large cohort of African American women (*N* = 40,403) that were stratified by age (< 60 and ≥ 60 years). Their findings also showed that correlations between factors were weaker in the older group, revealing the while the factor structure was confirmed, it was not invariant.

There have also been challenges to the four-factor model. Miller, Markides, and Black (1997) argue that a two-factor model was a more parsimonious model than the four-factor model for their minority population of elderly Mexican Americans (*N* = 2,866). The factors in their model include Depression Factor (16 items) and Well-being (4 items). Manson, Ackerson, Dick, Baron, and Fleming (1990) discuss the differences in parsimony between a two- and three-factor model in comparison to the four-factor model in a sample of American Indian boarding school students (*N* = 188; grades 9 – 12). They concluded that the two- and three- factor models were more desirable in terms of interpretability than the original factor structure. A three-factor solution was also shown to be more parsimonious in a sample of Chinese Americans (Ying, 1988); the factors that were retained were depression, positive affect, and interpersonal. Edman, Danko, Andrade, McArdle, Foster, and Glipa (1999) concluded that a two-factor solution fit reasonably well in a sample of Filipino-American adolescents (*N* = 243); factor 1 combined somatic-retardation, depressed affect, and interpersonal items, and factor 2 consisted of the remaining positive items.

The above examples highlight that symptomatology and the concept of depression differs cross-culturally, and thus Radloff's original factor structure may not be appropriate for varying subpopulations. The purpose of this paper is to investigate the factor structure of a specific subpopulation (impoverished African-American parents). While research has been done regarding cultural differences in depression, few have investigated impoverished populations; our population sample captures both culture and impoverishment. This is meaningful because items on a scale may not load as expected when the target population has specific traits that distinguish them from the overall population. In other words, these traits may influence the participants understanding and interpretation of the items, and therefore the factor structure may be altered. We investigate how the individual items load to their prescribed factors, and whether there is any disagreement with the results from this study compared to other published works. Based on the literature, the two-, four-, and higher-order factor solutions were the most parsimonious, and therefore we chose to compare these for our research purposes. The research questions for the present paper are:

RQ1: Is the factor structure of the CES-D for impoverished African Americans different from the general population?

RQ2: Which factor structure model best fits the data: two-factor, four-factor, or higher order model?

2. METHOD

2.1. Data and Instrument

The data come from the Mobile Youth and Poverty Study (MYPS), which consists of datasets related to poverty and adolescent risk in Mobile, AL. The test was administered annually to adolescents in impoverished neighborhoods between the years 1998 and 2011. Over 12,000 youths were enrolled in the MYS, and the age range is 10-18. In addition to the survey data, administrative databases have been accessed in order to provide further information about the participants including school records, court records, and housing records.

The particular set of questions from the MYPS used in the present study is the CES-D. The scale was developed and validated by Radloff (1977), and has gone through slight revisions over the past years. The CES-D scale is a short self-report scale designed to measure depressive symptomatology in the general population. Therefore, the scale should be a useful tool for epidemiologic studies of depression.

2.2. Participants

The participants (N = 1,020) were African American parents of children who are impoverished and at risk for adolescent behavioral issues. The age range for the sample population was 24 – 56 years, $\bar{x}$ = 36.2, SD = 4.1.

2.3. Analyses

For the statistical analysis, three different models were used. A preliminary exploratory factor analysis (EFA) was used to determine the factor structure for the dataset, and confirmatory factor analyses (CFA) were used to a) confirm the original factor structure of the CES-D that was established in Radloff (1977), b) compare the original four-factor structure to the two-factor structure based on the EFA results, and c) compare the original four-factor structure to a higher order factor model. A random sample of half of the participants ($n1$ = 510) were used in the EFA analysis, and the remaining participants ($n2$ = 510) were used in the CFA analyses. The statistical software package used for the analysis was SAS version 9.4. A parallel analysis was also conducted as an additional measure to check the dimensionality of the CES-D; this was performed in SPSS version 25.

2.4. Preliminary Exploratory Factor Analysis

In order to evaluate the factor structure for this particular population, a two- and four-factor solution was used for the EFA model. Upon inspection of the scree plot, it was determined that a two-factor solution was more efficient; the extraction sums of the squared loadings for the four- factor solution was 56%, while the two-factor solution retained 52.3%. The oblimin rotation method was used for the EFA analysis. The factor solution can be seen in Table 1. All factors loaded above .400 except appetite, showing that items are successfully loading to one of the factors. Lonely double loaded to both factors. Upon inspection of the Cronbach alpha levels, all items were retained in the solution. With a Kaiser-Meyer-Olkin measure of sampling value of .924, we can conclude that factor analysis is appropriate for the data. The determinate of the matrix is .202; therefore, the assumption that there are no linear dependencies in the data is met. Lastly, we did not detect outliers and participants were deleted using pairwise deletion methods.

Parallel Analysis. As an additional measure of dimensionality of the CES-D, a parallel analysis was conducted using the raw data. Based on the comparison of the raw data eigenvalues, mean, and percentile random data eigenvalues, the two-factor solution was the most parsimonious. The raw data eigenvalues were greater than the mean and percentile random data eigenvalues (1.443, 1.265, and 1.187, respectively).

Table 1. *Two-factor EFA model*

Variable	Factor 1	Factor 2
Dislike	.773	.119
Mind	.740	.101
Talk	.678	.123
Hopeful	.663	.071
Effort	.660	.046
Failure	.653	.054
Happy	.612	.085
Get going	.598	.044
Depressed	.566	.028
Fearful	.557	.009
Cry	.556	.001
Unfriendly	.507	-.036
Bothered	.498	.042
Sad	.486	-.006
Good	.477	.046
Lonely	.331	.653
Sleep	.229	.622
Blues	.010	.502
Enjoy	.005	.469
Appetite	.145	-.289

2.5. Two-factor Confirmatory Factor Analysis

Based on the two-factor EFA analysis, a CFA model was created. The results of the analysis are illustrated in Table 2. The fit statistics were consistent with those of a moderate-fit model, $\chi 2 = 1139.32$, $df = 168$, $p < .001$, RMSEA = 0.07, GFI = 0.89, and CFI = 0.85. Regarding the individual item loadings, appetite was the only nonsignificant item, with a loading of only .015. Blues and enjoy yielded relatively weak loadings, with values of .223 and .250, respectively. The remaining items had loadings above .300, and therefore are loading to the factors that were established in the preliminary EFA analysis.

Table 2. *Two-factor CFA model solution*

Variable	Factor 1	Factor 2
Bothered	.499*	
Good	.513*	
Mind	.801*	
Depressed	.460*	
Effort	.684*	
Hopeful	.641*	
Failure	.668*	
Fearful	.633*	
Happy	.612*	
Cry	.558*	
Sad	.489*	
Dislike	.835*	
Get going	.580*	
Unfriendly	.483*	
Lonely		.341*
Appetite		.015
Blues		.223*
Talk		.538*
Sleep		.496*
Enjoy		.250*

2.6. Four-factor Confirmatory Factor Analysis

Based on the original model by Radloff (1977), a four-factor model was specified. The results are shown in Table 3. The fit statistics were consistent with those of a moderate-fit mode, $\chi2$ = 1227.77, *df* = 164, *p* < .001, RMSEA = 0.08, GFI = 0.88, and CFI = 0.83. Factor 1 had a weak loading of .094 for blues. Enjoy had a weak factor loading of .093 to factor 2. Appetite had a nonsignificant weak loading of .108 to factor 3. The remaining items loaded on a factor at or above .300, so these items on the scale are loading on the factors they were designed to.

Table 3. *Four-factor CFA model solution*

Variable	Factor 1	Factor 2	Factor 3	Factor 4
Blues	.094*			
Depressed	.455*			
Failure	.645*			
Fearful	.607*			
Lonely	.448*			
Cry	.553*			
Sad	.471*			
Good		.520*		
Hopeful		.657*		
Happy		.632*		
Enjoy		.093*		
Bothered			.491*	
Appetite			.108*	
Talk			.609*	
Mind			.787*	
Sleep			.328*	
Get going			.579*	
Effort			.669*	
Dislike				.855*
Unfriendly				.486*

2.7. Higher Order Confirmatory Factor Analysis

Upon inspection of the covariances among the exogenous variables, it was determined that a higher order model be specified; previous literature regarding the CES-D shows that the four factors developed are assumed to be correlated (Radloff, 1977), adding further justification for a higher order solution. The covariances among the exogenous variables are displayed in Table 4, and the covariance matrix for the four-factor model is shown in Table 5. The fit statistics were consistent with those of a moderate-fit mode, $\chi 2$ = 1228.86, df = 166, p < .001, RMSEA = 0.08, GFI = 0.88, and CFI = 0.83. Appetite and blues had weak loadings on factor 1 (.102 and .066, respectively). Factor 3 yielded a nonsignificant weak factor loading of .056 for enjoy. The remaining items loaded on a factor at or above .300, so these items on the scale are loading on the factors they were designed to.

Table 4. *Covariances among exogenous variables*

	Factor 1	Factor 2	Factor 3	Factor 4
Factor 1	1.000			
Factor 2	.973*	1.000		
Factor 3	1.063*	.993*	1.000	
Factor 4	1.002*	.935*	.999*	1.000

* *statistically significant*

Table 5. *Higher order CFA model solution*

Variable	Factor 1	Factor 2	Factor 3	Factor 4	General Factor
Blues	.081*				
Depressed	.568*				
Failure	.633*				
Fearful	.533*				
Lonely	.414*				
Cry	.535*				
Sad	.463*				
Good		.482*			
Hopeful		.694*			
Happy		.635*			
Enjoy		.079*			
Bothered			.491*		
Appetite			.091*		
Talk			.680*		
Mind			.738*		
Sleep			.310*		
Get going			.591*		
Effort			.678*		
Dislike				.802*	
Unfriendly				.494*	
Factor 1					1.027*
Factor 2					.957*
Factor 3					1.034*
Factor 4					.969*

4. DISCUSSION

In terms of first research question, it appears as the original factor structure of the CES-D is slightly different for this sample population of impoverished African Americans than the general population. The overall model fit indices were adequate, even though there were slight departures from the expected loadings of certain variables. While the original factor structure moderately fit the data, it is worth exploring other factor models that may better fit the current population sample.

In terms of the second research question, deciding on a best fitting model is not as obvious as we would hope. When considering the model fit indices, the two-, four-, and higher- order factor solutions all produce moderately fitting models. However, upon closer inspection of the item loadings, in the four-factor and higher order models, it appears that there are a few items that not contributing to the CES-D the way they were originally designed. Across these models, appetite, blues, and enjoy were either nonsignificant or produced weak loadings to their factors. This not consistent with finding in previous research that looks at the factor structure of the CES-D in minority populations (Roberts, 1980; Williams, 2007). However, in the two-factor solution, appetite was the only item to have a weak loading (.015). Blues and enjoy had adequate loading values (.223 and .250, respectively). We argue that the two-factor model best fits the current data because of the higher overall factor loadings and is the most parsimonious model.

These findings are consistent with previous research that shows a two-factor solution appears to fit the data better with minority populations (e.g. Manson et. al., 1990 and Edman et. al., 1999). It appears as though this population sample of impoverished African-American parents are interpreting the items of the CES-D differently than the population the CES-D was originally developed for and are instead interpreting closely to other minority populations such as American-Indian and Filipino-American. Our recommendation is for researchers to consider using a two-factor solution of the CES-D with African samples from lower socioeconomic backgrounds. However, further research is needed to fully understand the relationship between impoverished African-Americans and other minority groups. Researchers should also continue investigating the effects socioeconomic status has on depression, as well as how the factor structure of the CES-D (and other depression scales, for that matter) are impacted.

When deciding on a best fitting model, the actual loadings need to be considered as well. The loadings as a whole are higher in the two-factor model than the four-factor or higher-order model. The two-factor model produced better loadings, in comparison to the other models, in the following items: bothered, appetite, blues, mind, failure, fearful, sleep, enjoy, cry, and sad. Five of these items are originally on factor 1 (depressed affect), and four originally loaded to factor 3 (somatic and retarded activity) in the four-factor solution established by Radloff (1977). We argue that the two-factor solution is more parsimonious and can better describe the items on the scale related to depression affect and somatic and retarded activity than the four-factor model for this minority population. The factor structure as illustrated in Table 2 is the solution we encourage researhers to considere when evaluating depression via the CES-D in similar samples.

One of most difficult tasks given to a measurement researcher is naming the factors developed using CFA. The relation between the items in factor 1 and factor 2 seems to be feeling versus action, respectively. Items such as bothered, depressed, hopeful, fearful and unfriendly appear to be focused on an individual’s feelings or attitude, while appetite, talk, and sleep are more related to individuals’ actions. Therefore, we would tentatively name the factors feeling and action.

Regarding the items that failed to adequately load to their respective factors in the four-factor and higher-order models, we can offer a few possibilities to why these items are not loading to the factors they were designed to. Perhaps the social context of the impoverished African Americans is affecting their interpretation of the items on the CES-D. Their living environment may not be conducive to experiencing reliable levels of enjoyment. High crime rates, social injustice, and broken families may interfere with their interpretations of blues and enjoy. (Cutrona, et al., 2005). Other studies have found that perceived discrimination and social injustice and poorer mental health could be associated (Brown et. al., 2000). In a longitudinal analysis, Schulz et. al. (2006) found that was a positive relationship between a change in perceived discrimination over time and a change in symptoms of depression, further highlighting the impact that perceived discrimination has on mental health.

Another possible covariate to explore is socioeconomic status (SES) in relation to an individual's depression. Few articles have explored the direct effects of SES on depression, though this is evidence to support the claim that SES is related to depression. Previous research has shown that individuals with higher education can more successfully delay increased levels of depression over time (Miech & Shanahan, 2000). Additionally, an individual from a high SES may have an interpretation of enjoyment and depression that is vastly different than someone from a low SES (Schnittker, 2008). CFA models such as bifactor models could be used to explore if there is an underlying affect that is related to the factors in the CES-D.

It is worth noting that a possible methodological limitation to our study was the lack of control of age in our population. The range of age in our population was large, and thus it is possible for age to also be an important covariate to explore. However, with a sample mean age of 36 years and *SD* of 4.2, we believe that age was not a large enough factor to detrack from our findings. Future studies could investigate age to determine if it is a significant covariate of depression.

ORCID

Mitchell Porter https://orcid.org/0000-0001-8589-4880
Youn-Jeng Choi https://orcid.org/0000-0001-9803-2681
Sara Tomek https://orcid.org/0000-0003-0705-3087

5. REFERENCES

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*(6), 561-571.

Brown, T. N., Williams, D. R., Jackson, J. S., Neighbors, H. W., Torres, M., Sellers, S. L., & Brown, K. T. (2000). "Being black and feeling blue": The mental health consequences of racial discrimination. *Race and Society*, *2*(2), 117-131.

Campbell, L.L. (2007). CES-D four-factor structure is confirmed, but not invariant, in a large cohort of African American women. *Psychiatry Research*, *150*(2), 173-180.

Cosco, T. D., Prina, M., Stubbs, B., & Wu, Y. T. (2017). Reliability and validity of the Center for Epidemiologic Studies Depression Scale in a population-based cohort of middle-aged US adults. *Journal of Nursing Measurement*, *25*(3), 476-485.

Clark., V.A., Aneshensel, C.S., Freriches, R.R., & Morgan., T.M. (1981). Analysis of effects of sex and age in response to items on the CES-D scale. *Psychiatry Research*, *5*(2), 171-181.

Cutrona, C. E., Russell, D. W., Brown, P. A., Clark, L. A., Hessling, R. M., & Gardner, K. A. (2005). Neighborhood context, personality, and stressful life events as predictors of depression among African American women. *Journal of Abnormal Psychology*, *114*(1), 3.

Edman, J. L., Danko, G. P., Andrade, N., McArdle, J. J., Foster, J., & Glipa, J. (1999). Factor structure of the CES-D (Center for Epidemiologic Studies Depression scale) among

Filipino-American adolescents. *Social Psychiatry and Psychiatric Epidemiology*, *34*(4), 211-215.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, *23*(1), 56.

Hertzog, C., Alstine, J.V., Usala, P.D., Hultsch, D.F., & Dixon, R. 1990. Measurement Properties of the Center for Epidemiological Studies Depression Scale (CES-D) in Older Populations. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*(1), 64-72.

Husaini, B.A., Neff, J.A., Harrington, J.B., Hughes, M.D., & Stone, R.H. (1980). Depression in rural communities: Validating the CES-D scale. *Journal of Community Psychology*, *8*, 20-27.

Knight, R. G., Williams, S., McGee, R., & Olaman, S. (1997). Psychometric properties of the Centre for Epidemiologic Studies Depression Scale (CES-D) in a sample of women in middle life. *Behaviour research and therapy*, *35*(4), 373-380.

Lewinsohn, P. M., Seeley, J. R., Roberts, R. E., & Allen, N. B. (1997). Center for Epidemiologic Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. *Psychology and Aging*, *12*, 277-287. doi:10.1037/0882-7974.12.2.277

Manson, S.M., Ackerson, L.M., Dick, R.W., Baron, A.E., and Fleming, C.M. (1990). Depressive symptoms among American Indian adolescents: Psychometric characteristics of the Center for Epidemiologic Studies Depression Scale (CES-D). *Psychological Assessment*, *2*(3), 231-237.

Miech, R. A., & Shanahan, M. J. (2000). Socioeconomic status and depression over the life course. *Journal of health and social behavior*, *41*(2), 162-176.

Miller, T.Q., Markides, K.S., & Black, S.A. (1997). The factor structure of the CES-D in two surveys of elderly Mexican Americans. *Journal of Gerontology: Social Sciences*, *52B*(5), S259 – S269.

Montgomery, S. A., & Åsberg, M. A. R. I. E. (1979). A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry*, *134*(4), 382-389.

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385-401.

Roberts. R.E. (1980). Reliability of the CES-D scale in different contexts. *Psychiatry Research*, *2*, 125-134.

Santor, D.A. & Coyne, J.C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. *Psychological Assessment*, *9*(3), 233-243.

Schnittker, J. (2008). Happiness and success: Genes, families, and the psychological effects of socioeconomic position and social support. *American Journal of Sociology*, *114*(S1), S233-S259.

Schulz, A. J., Gravlee, C. C., Williams, D. R., Israel, B. A., Mentz, G., & Rowe, Z. (2006). Discrimination, symptoms of depression, and self-rated health among African American women in Detroit: results from a longitudinal analysis. *American Journal of Public Health*, *96*(7), 1265-1270.

Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology*, *62*(1), 123-146.

Sommel, M., Given, B.A., Given, C.W., Kalaian, H.A., Schulz, R., & McCorkle, R. (1993). Gender bias in the measurement properties of the center for epidemiologic studies depression scale (CES-D). *Psychiatry Research*, *49*(3), 239-250

Stroup-Benham, C. A., Lawrence, R. H., & Trevifio, F. M. (1992). CES-D factor structure among Mexican American and Puerto Rican women from single-and couple-headed households. *Hispanic Journal of Behavioral Sciences*, *14*(3), 310-326.

Williams, C.D., Taylor, T.R., Makambi, K., Harrell, J., Palmer, J.R., Rosenberg, L., & Adams-

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., & Leirer, V. O. (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of Psychiatric Research*, *17*(1), 37-49.

Ying, Y. W. (1988). Depressive symptomatology among Chinese-Americans as measured by the CES-D. *Journal of Clinical Psychology*, *44*(5), 739-746.

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 602–616
https://dx.doi.org/10.21449/ijate.554212

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

Investigating Differential Item Functioning of Ankara University Examination for Foreign Students by Recursive Partitioning Analysis in the Rasch Model

Özge Altıntaş [1,*], **Ömer Kutlu** [1]

[1]Department of Educational Measurement and Evaluation, Ankara University, Turkey

ARTICLE HISTORY

Received: 15 April 2019

Revised: 26 July 2019

Accepted: 10 November 2019

KEYWORDS

Differential Item Functioning,
Rrecursive Partitioning Analysis,
Rasch Tree Method,
Examinations for International Students

Abstract: This study aims to determine whether items in the Ankara University Examination for Foreign Students Basic Learning Skills Test function differently according to country and gender using the Recursive Partitioning Analysis in the Rasch Model. The variables used in the recursive partitioning of the data are country and gender. The population of the study is composed of 2476 individuals. Since the study includes comparisons across countries, the country is accepted as a criterion in determining the sample group. Thus, the sample of the study consists of 615 individuals selected from Azerbaijan, Bulgaria, and Syria. To investigate differential item functioning (DIF) of the items of the test, the Rasch tree method was used. As a result of the analysis, DIF has been detected in 16 items at the 0.001 significance level. However, these items have been identified to have similar difficulty parameters in all countries. Finally, items have not shown DIF according to gender.

1. INTRODUCTION

The constant social, economic, political and cultural changes have encouraged societies to explain the world in which they live based on the sovereign power of knowledge. It is possible to see the products of the mind in the transition to agriculture, emergence of cities, and birth of urban-state civilization. Interest, curiosity and needs have led our ancestors to learn about the natural world in terms of food, heating and protection, and use it for their own benefit. Since the 1600s, resources that help disseminate information; e.g., the printing press, and those that increase production, such as steam, coal, and electricity have entered the life of societies (Asimov, 2004).

The effort to understand human and social life in the new century has further revealed the value of scientific knowledge and scientific research. Scientific information has contributed to the rapid development of information technologies. With the emergence of new technologies,

CONTACT: Özge Altıntaş ✉ oaltintas@ankara.edu.tr Department of Educational Measurement and Evaluation, Ankara University, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

social life has undergone a change and has been reshaped, thus the need for continuous regeneration of information required to sustain life has grown.

Since scientific knowledge is a shared value, incomplete, erroneous or even inaccurate information has a negative effect on the changes of societies. Toffler (1980) stated that change is not linear; it is forward, backward or lateral. Transformation of change into social development; i.e., forward change, depends on the production of accurate knowledge, which requires the application of principles of reason and logic to understand and explain natural and social events. This is closely related to raising individuals that value scientific knowledge, seek information, and know how to obtain it.

Today, many societies consider that economic, cultural and political development is only possible through raising qualified and educated individuals. Therefore, they attach importance to all the educational stages, especially preschool education, and the construction educational institutions in accordance with the requirements of the era. The increase in the number of educated individuals leads to an increase in individuals who want to receive higher education. Thus, societies not only open new higher education institutions and programs but also cooperate with foreign educational institutions to ensure that their citizens receive a better quality of education (Altıntaş, 2016).

Many graduates of secondary education programs want to study in various higher education programs of universities outside their country in order to obtain a higher quality university education. Those countries that accept international students try to determine the application requirements for those candidates. Such education programs in Western countries that accept a large number of overseas students require internationally recognized high school diploma degrees and/or scores from internationally valid tests, such as the Scholastic Aptitude/Assessment Test (SAT), the American College Testing (ACT), and the Thinking Skills Assessment (TSA). These tests are prepared in English, a language which is widely used in the world, and the test items include verbal, numerical and formal expressions. The items mainly measure high-level thinking processes, such as reasoning, critical thinking, problem solving, and abstract thinking (Zwick & Sklar, 2005). In terms of the measured characteristics, these tests aim to assess high-level mental processes; e.g., using, interpreting and generalizing knowledge, making differentiations, and establishing and evaluating relations between different components (Kutlu & Karakaya, 2007).

Selection and placement tests implemented in educational processes are used in the transition of students from the present learning step to a higher educational stage and in deciding whether the student can move to the upper level (Cronbach, 1990). The selection and placement tests currently used in Turkey differ from those implemented from 1961 to 1980 in that they aim to measure the academic ability of the students and consist of items based on secondary education programs (Oral, 1985). These tests measure the students' ability to use the basic knowledge and skills acquired in school programs, and in 1981, this new approach was also adopted for the development of Examination for Foreign Students (YOS) tests, used in the selection and placement of foreign students.

In 2010, the Turkish Council of Higher Education (YOK) examined the process of foreign student selection with a view to increasing the competitive power of Turkey in the international arena. In parallel to the studies concerning student placement in the higher education system in the country, beginning from the 2010-2011 academic year, YOK abandoned the YOS system and decided that universities were to determine the principles and conditions to be applied for the admission of foreign students and receive YOK's approval before implementing them (YOK, 2013).

In accordance with this decision, universities began to determine their own programs and quota for foreign students with the approval of YOK. Within the framework of these principles, through the decree of the academic senate, the Rectorate of Ankara University accepted SAT 1, Abitur, the International Baccalaureate Diploma, and the scores in Ankara University Examination for Foreign Students (AYOS) as the criteria to be considered in the selection and placement of foreign students to study in the university (Ankara University, 2013). Since the scores obtained from the tests used in these exams differ, the scores of diplomas are converted by the university's Student Affairs Office to comply with the scoring of AYOS. However, the lack of equivalence in the scoring of placement tests constitutes a measurement problem (Altıntaş, 2016).

The Student Selection and Placement Center (OSYM) tried to maintain the goal of selecting and placing students in higher education in the 1980s, at which time this center was responsible for determining the framework of foreign student examinations. The items included in the tests that aim to select and place foreign students in higher education programs in Turkey require the student to use thinking processes, such as comprehension, application, and analysis (Toker, 1997; USYM, 1978, 1980a, 1980b).

These items are based on the relation between figures, numbers, and letters, which are independent of language. In addition, they are associated with mental processes; e.g., analytical thinking, reasoning, and abstract thinking that develop in individuals over a long period of time. The reason why such tests are developed in a language-independent manner; i.e., containing very limited verbal language and mostly utilize figures, numbers and letters, is that word use, relationships between words, and verbal instructions are not suitable for those individuals who do not have an adequate knowledge of the target language (Resing, 2005).

Since 2011, AYOS has been conducted in accordance with the aims of YOS tests and the general purposes of developing student selection and placement tests in higher education. In this exam, the Basic Learning Skills Test (TOBT) consisting of 100 items is used. The items are prepared independently of verbal expression, language, and the content of the curricula of the schools. The first 60 items of test are based on the relationships between shapes, numbers, and letters, the items measuring psychological properties, such as analytical thinking, reasoning, abstract and spatial thinking; the remaining 40 items consist of the items that measure numerical thinking skills that require the use of mathematics and geometry information (ANKUDEM, 2011). In other words, the test is a "non-verbal or verbal neutral" measurement tool. This is mostly because the foreign students preferring to study in higher education programs in Turkey are coming from different cultures and therefore, they do not know Turkish or another foreign language well.

The first 60 items in TOBT mainly aim to measure students' analytical thinking, abstract thinking, and reasoning. Analytical thinking is the process of breaking things down into their constituent components in order to understand the whole and examining the relationships between these components. Reasoning refers to the process of making inferences and reaching a conclusion based on the information given (Bruner, 1957 as cited in Lohman & Lakin, 2011). The remaining 40 items measuring students' ability to use numerical skills based on basic mathematics and geometry predominantly require the individuals to establish connections based on shapes, numbers and letters, make logical inferences, and engage in abstract thinking and reasoning.

Since the beginning of the 1900s, many researchers have attempted to measure skills, such as analytical thinking, reasoning, and abstract thinking through intelligence tests based on the relationships between shapes, numbers and letters. Looking at the process related to psychological measures, the tools used to measure the mental abilities of individuals are now known as academic aptitude tests. These tests also contribute to the estimation of school achievement by identifying individuals' abilities (Anastasi, 1979). Walsh and Betz (1995) stated that aptitude tests used in education can also help predict future educational achievement. The high correlation between the scores obtained from the academic aptitude tests and the academic achievement scores indicates that the students with higher aptitude scores may have higher school grades than those with lower aptitude scores (Sternberg, 1997).

Certain psychometric properties are sought in the tests designed to measure academic aptitude, one of which is the tests of having predictive power. Thus, one of the most important features of an academic aptitude test prepared for student selection and placement is predictive validity. The predictive power of a test means that the feature measured by the test in a given period is related to a particular feature in the future. Prediction is a matter of making estimations, and predictive validity is often used in the selection of tests for educational placement and recruitment for employment (Turgut & Baykul, 1992).

Tests developed for student selection and placement should be able to predict achievement scores to be obtained from future tests based on the scores in the currently applied tests (Thornell & McCoy, 1985). One of the factors that influence the predictive power of a test is the items representing the psychological characteristics that are measured. For a test to have high predictive power, the items contained should measure the mental characteristics that develop in individuals over a long period, rather than psychological features that develop in a shorter time (Anastasi & Urbina, 1997; Cronbach, 1990).

Psychological measurement tools are affected by demographic properties, such as age and gender, and cultural properties; e.g., ethnicity and country, whether developed in verbal or non-verbal language (Messick, 1989). This situation makes the accuracy of the decisions given based on the scores obtained from the measurement tools questionable. Especially in the selection and placement tests, the items which are important for students need to have equal response possibility for those who have the same ability level. Otherwise, individuals with the same ability level will be advantageous or disadvantageous compared to each other.

An essential part of the test development and item preparation process in the education and psychology fields is to divide the tests into different groups and determine whether the measurement results are the same for each group. In particular, the items included in selection and placement tests, which have an important impact on the lives of individuals, must create the possibility of an equal response from individuals at the same level of ability and skills. In other words, the test items should not be biased toward certain subgroups. Otherwise, individuals with the same skill level taking the test will gain an advantage or be disadvantaged in relation to each other. Since AYOS is developed for selection and placement purposes, it is necessary to examine the psychometric properties of these tests to ensure that such bias is not present.

Karakaya and Kutlu (2012) state that the determination of bias of the items in the tests is one of the important studies to increase the validity and reliability of the test. Therefore, undertaking an investigation of item-level bias is very important in the test development process, especially in the process of item writing, in establishing the preliminary evidence on the validity.

The concept of validity is addressed and defined from different perspectives in the literature due to the variety of validation methods. According to Anastasi (1979), validity refers to what a measurement tool measures and the extent of which the measurement process is undertaken correctly. A measurement tool applies only to a specific purpose under certain conditions; it cannot be asserted that the same measurement instrument applies to other purposes or conditions. Messick (1989) defined validity as a degree of all assessments supporting the accuracy and appropriateness of implications related to the scores of a measurement tool or other measurement cases based on theoretical information and empirical evidence. In this respect, validity is not only a feature of measurement or evaluation; rather, it is concerned with the meaning of the scores obtained from the measurement tool. Linn and Gronlund (2000), including all these definitions, described validity as the most fundamental and important factor in determining the accuracy and appropriateness of implications and judgments about the results obtained from measurement tools.

The issue of validity, discussed in detail in the Measurement Standards in Education and Psychology, prepared in 2014 by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), has found wide coverage in *Standard 1.0* to *Standard 1.25*. In the same source, validity is defined as the degree of evidence and theories supporting the interpretability of the scores obtained from tests. According to this definition, validity describes an ongoing process, rather than merely results. In this process, additional information is always available to achieve a better understanding of the implications of the test. Making inferences about validity is similar to undertaking scientific inferences. Therefore, validity studies are conducted by presenting supporting information about test scores. A prerequisite for this, as emphasized in *Standard 1.1*, is to confirm that the construct(s) intended to be measured in a test are clearly defined and are not extensively linked to other constructs.

The *Test Validation* section of *Educational Measurement*, written by Cronbach (1971) and edited by Thorndike, emphasizes the importance of psychological constructs in educational measurement. According this source, whenever a test developer[†] asks, "What does this measurement tool really measure?", s/he wants to obtain information about construct validity. In this respect, a test being able to measure the construct of interest despite the presence of disrupting effects emerges as a condition to which testers pay great attention.

Whether the test is appropriate or powerful enough to measure the intended construct is determined by obtaining evidence on construct validity because the power of a test to measure a construct is an important indicator of the extent to which the test serves its purpose (Linn, 1989). Therefore, considering that the items in a test are also developed in accordance with the purpose of that test, it can be assumed that investigation of construct validity begins with the test development process.

Cronbach and Meehl (1955) emphasized that construct validity was important for tests that measure many psychological features, such as interest, attitude, ability, and success. By the beginning of 1900s, studies starting with Charles Edward Spearman, led psychometrists to investigate the real value of the observed characteristics. Since then, the importance of the reliability and validity values of measurement instruments has been acknowledged, and the related coefficients have become the most important criteria for these tools. Achieving a certain reliability and validity value for the scores obtained from measurement tools is considered as a requirement but not sufficient alone for the measured characteristics. This understanding has resulted in deepening the studies on the structural features of measurement instruments. For example, the tests used in education have been enriched in terms of psychometric features measured, and national standard achievement tests and more recently large-scale tests that can be used at the international level have been developed. The widespread use of tests has raised the question of whether the items contained in these tests can measure the intended constructs independently of individuals and their associated groups (Cohen et al., 1988; Crocker & Algina, 1986).

After 1950s, many researchers began investigating the reasons for the conflicting results obtained from intelligence tests. The 1960s mark the discussion of issues related to the fair use

[†] The word 'educator' is used in the original source. However, 'test developer' was used here in accordance with the context.

of tests and item bias, which was mainly a result of the human rights movement initiated in the United States of America in 1964. This movement led to the signing and enforcement of certain laws on equality and equal opportunities. The fundamental changes brought about by the human rights movement also attracted the attention of test developers to the use of tests that may have an adverse impact[‡] on recruitment and education (Osterlind & Everson, 2009). After the 1970s, psychometrists concentrated their research on this issue of 'bias', for the development of tools that would make more objective and sensitive measurements (Reynolds & Suzuki, 2013).

In the following years, the related studies were not limited to a specific culture, but were deepened by extensive research on cross-cultural comparisons. However, during these investigations that focused on bias in tests developed to measure psychological features, the items included in tests, and individuals (gender, age, etc.) or groups (culture, ethnic origin, etc.) responding to these items, researchers commonly faced four important problems and limitations: (1) lack of a consensus on the definition of psychological constructs or characteristics to be measured, (2) failure to select a sample that represents the groups to be compared, (3) inability to standardize the conditions of the test application (absence of standardization), and (4) lack of rules regarding the translation, adaptation and scoring of test booklets (Hambleton, 2002).

Psychological measurement tools are developed to obtain information about the psychological characteristics of individuals living in a particular culture. Thus, a measurement tool developed in a culture possesses features specific to that culture (Öner, 1987). Culture is a very important factor in the test development process since it can affect the scores obtained from tests and the psychometric characteristics related to these scores (Hambleton, Merenda, & Charles, 2005). Wicherts (2007) noted that the results of the measurement might differ according to individual characteristics, but it would be wrong to attribute these differences to individual characteristics alone since they might also result from the measurement tool. For example, assuming that a girl and a boy have a similar level of knowledge in mathematics, if there is a systematic difference between the scores of these students in a mathematics test developed to measure the related construct, it can be stated that the test has a gender bias.

The first known studies on bias date back to the 1900s, when it was determined that the scores of socioeconomically disadvantaged children who had taken the intelligence test developed by Alfred Binet were related to what they had learned at home or school, rather than the mental characteristics of the individuals, which led to the removal of certain items from this test (Camilli & Shepard, 1994).

According to Crocker and Algina (1986), bias research has two main objectives. First refers to whether the test scores are affected by different variance sources in different subgroups taking the test, and second is whether the test scores are affected by the same variance sources for all subgroups. If the test scores are judged to be affected by the same variance sources in all subgroups, it should also be investigated whether there are unrelated sources that provide unfair advantage to certain subgroups.

There are two basic statistical approaches to the investigation: external methods, in which an external measure independent of the test is used, and internal methods, in which psychometric features of the items included in the test are used as criteria. In external methods, analysis of bias is undertaken by comparing the differences in averages of the total scores obtained from a test for different subgroups to those obtained from a different test considered to measure the same construct. In cases where an external measure is not available, internal methods can be used to determine bias by examining the psychometric features of the items included in the test (the total test score obtained from the items and skill level) (Shepard, Camilli, & Averill, 1981).

[‡] A term used to refer to evidence supporting unlawful discrimination claims.

However, item analysis methods, such as item-total score correlations performed in accordance with the Classical Test Theory (CTT) or variance analysis that compares the items and total test scores in a similar manner do not provide sufficient information about bias based on the averages of groups and differences between these averages. Bias investigations undertaken using such methods may have deficiencies since the psychometric features of items and the total scores obtained from the test are affected by the skill distribution of the sample in CTT. Therefore, the differences in individuals' test performance or average scores in test items should not be interpreted as evidence of a bias in direct comparison groups.

In the third section of the Measurement Standards in Education and Psychology[§], AERA, APA and NCME (2014) present 20 standards related to the fairness of the test scores; i.e., they should be free from bias. In all these standards, it is emphasized that the test or item scores should have the same meaning for all individuals (within subgroups) that have taken the test, and that the test scores should be comparable. *Standards 3.2* and *3.13* explain in detail the necessity to take into account the different characteristics of the test respondents, such as language and culture, in accordance with the purpose of the test. If there are thought to be differences in the test and item performance of respondents in terms of the measured characteristic according to ethnicity, language, culture, gender and age groups, these differences should be investigated in as much depth as possible.

In other words, the tests to be administered to the individuals from different groups and the items to be included in these tests should be designed in such a way as to reduce the situations that may lead to bias. Therefore, the evidence related to whether or not the measurement instrument investigated in this study, TOBT, caused a bias for/against individuals in different subgroups was examined by investigating whether the function of the TOBT items differed between these individuals. Accordingly, the aim of this research was to investigate whether the items contained in the Basic Learning Skills Test in the AYOS-TOBT 2017 showed differential item functioning (DIF) according to three countries (Azerbaijan, Bulgaria, and Syria) and gender.

2. METHOD

2.1. Research Design

This research is a survey type taken from descriptive research models. The study uses a descriptive research model, since it aims to investigate whether the items of AYOS-TOBT show DIF in terms of country and gender variables and describe the current situation (Karasar, 2015). Descriptive research is describing and interpreting the factors that are the subjects of the study; however, this goes beyond gathering and classifying the data. Research process also includes collecting, classifying, describing, analyzing and inferring results results from the data (Best, 1970).

2.2. Population and Sample

The population of the study is 2476 individuals (N_{female} = 1184 approx. 48%, N_{male} = 1292 approx. 52%) from 75 countries who took AYOS 2017. Based on the purpose of the research, a purposive sampling method was used to select the sample in order to conduct an in-depth research and obtain rich information. In this study, criterion sampling was used from purposive sampling methods (Büyüköztürk et al., 2015). Since the study includes comparisons across different countries, culture was accepted as a criterion in determining the sample group. Thus, it was represented in both the number of students and different cultures, and the sample of the study consisted of 615 individuals selected from Azerbaijan, Bulgaria, and Syria. The distribution of the sample according to country and gender is given in Table 1.

[§] Fairness in Testing

Table 1. Distribution of sample by country and gender

Gender / Country	Female		Male		Total	
	n	%	n	%	n	%
Azerbaijan	51	38.35	82	61.65	133	21.63
Bulgaria	125	58.69	88	41.31	213	34.63
Syria	107	39.78	162	60.22	269	43.74
Total	283	46.02	332	53.98	615	100.00

The numbers of individuals taking the exam from Bulgaria (213, approx. 35%) and Syria (269, approx. 44%) were close to each other than Azerbaijan (133, approx. 22%). Besides, the number of individuals in the sample was close to each other in terms of gender (n_{female} = 283 approx. 46%, n_{male} = 332 approx. 54%) (Table 1).

2.3. Data collection

The research data was composed of the students' responses to AYOS-TOBT 2017, simultaneously implemented in three different exam centers located in Ankara/Turkey, Cologne/Germany and Baku/Azerbaijan in a single session. The responses obtained from two different booklets (A and B) were reordered according to Booklet A, and the data were prepared for analysis by converting it to the 1-0 scoring matrix and merging.

In this study, in order to limit the number of items examined, the first 60 items of the test were selected since they were considered to be more similar to the characteristics measured.

2.4. Data Analysis

In this study, bias analysis performed at the item level in terms of different subgroups was undertaken using a DIF analysis within the scope of the Rasch Model. This approach of model-based recursive partitioning (MBRP), proposed by Zeileis, Hothorn, and Hornik (2008), includes tests for both predefined groups and all possible groups without complicating the interpretation process. This allows for the determination of parameter imbalances. Similar to implicit class or mixed models, the main idea underlying this approach is based on identifying the groups in which the model parameters are differentiated is the sequential testing of all groups by investigating all possible sources that may cause DIF. In recursive partitioning, groups are defined not by an implicit factor as in implicit class models, but through the combinations of the observed common variables, based on an intuitive approach. Thus, MBRP offers intuitive, yet easy-to-interpret alternatives to implicit class or mixed models.

The Rasch tree model, a very new method for determining DIF, is based on MBRP, in which tests for structural change adapted from econometrics are used. MBRP is highly correlated with classification and regression tree methods, in which a common variable field is recursively partitioned to determine the group of a categorical or continuous response variable with different values. MBRP has a semi-parametric approach including the parameters of a parametric model varying between groups instead of values for a single response variable. Such parameters may be those of the Rasch model, which vary between groups or constant and slope parameters of a linear regression model (Strobl, Kopf, & Zeileis, 2015).

In MBRP analysis, the purpose is to divide the data matrix into subgroups (classes) with a homogenous structure. Each of these subgroups is called a node. Subgroups are primarily defined by common variables (such as age and gender). Then, these nodes are broken down as in classification and regression trees, like the branches of a tree, until they have a homogenous structure within. This is known as a Rasch tree, and the branches of this tree contain critical values (leaves) for common variables. This process continues until each node has the lowest variance value and the variance between the nodes is the highest (Kopf, Augustin, & Strobl, 2010).

In this regard, a Rasch tree with each leaf containing a node associated with a suitable model (e.g., maximum likelihood model or linear regression model) is created to achieve model-data fitness. Here, the basic idea is that each node is associated with a single model. Below is the successive steps (algorithm) used to create a Rasch tree (Strobl, Kopf, & Zeileis, 2015):

1. Estimate the item parameters jointly for all subjects in the current sample, starting with the full sample.
2. Assess the stability of the item parameters with respect to each available covariate.
3. If there is significant instability, split the sample along the covariate with the strongest instability and at the cutpoint leading to the highest improvement of the model fit.
4. [Repeat Steps 1-3 recursively in the resulting sub-samples until there are no more significant instabilities (or the sub-sample becomes too small)].

The Rasch tree employs the model-based recursive partitioning algorithm to detect groups that display different item parameters in the Rasch model (Kopf, 2013). This analysis was performed using the 0.12-1 version of the add-on package PsychoTree (Zeileis et al., 2011) in R program, which is open source statistical software (R Core Team, 2013). The Psychotree package was used to identify the items that showed DIF. The data was analyzed using the Rasch tree method (RTM) included in Psychotree, and the items showing DIF according to country and gender were determined. The significance level for the determination criteria of DIF, usually set to 5%, serves as the most important stopping criterion (Strobl, Kopf, & Zeileis, 2015). In this study, in addition to the 0.05 level of significance, a lower value, 0.01, was used as the DIF criterion in RTM.

3. FINDINGS and DISCUSSION

Table 2 shows the parameter instability tests conducted to determine whether the TOBT items indicated DIF at the level of significance according to the country and gender.

Table 2. Parameter instability tests: Test statistics and their corresponding p values by country and gender

Cov.	Par. Inst.	Node 1	Node 2
Country	statistic	.000	.000
	p value	.000*	.000*
Gender	statistic	90.2769556	85.3219230
	p value	0.2794151	0.4997366

*$p < .001$

As shown in Table 2, the country was accepted as a covariant, since the instability statistics obtained according to the country variable was significant. Gender was not considered a covariant since the value obtained for the gender variable was not significant. The Rasch tree showing this situation is given in Figure 1.

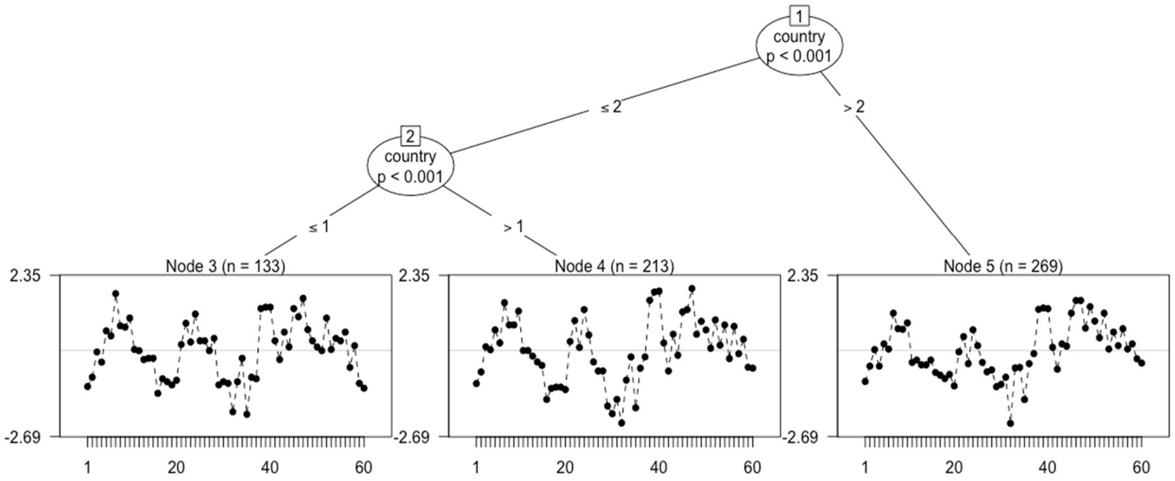


Figure 1. Rasch tree by country

As shown in Figure 1, the estimates of the difficulty parameters of 60 items in TOBT were between 2.35 and -2.69. The high level of these values means that the items were difficult while those with low values were easy (Strobl, Kopf, & Zeileis, 2015). Accordingly, it can be stated that the difficulty levels of the items were close to the average value of zero. Figure 1 also includes the significance values for each partition. Hence, there was a variation in terms of the countries in which the partition took place.

The Rasch tree obtained also provides information on which countries the difference appears. Thus, Syria differs according to the other two countries (Azerbaijan and Bulgaria) in the first partition and Azerbaijan differs from Bulgaria according to the second partition. When the Rasch tree is examined in detail, the 16 items (3, 4, 5, 6, 11, 12, 21, 23, 26, 27, 28, 44, 51, 53, 55, 57) in the TOBT included DIF in terms of countries. Distribution of difficulty parametres of the 16 items showing DIF is shown comparatively in Table 3.

Table 3. Distribution of difficulty parameters of items showing DIF according to the countries

Item no.	Countries		
	Azerbaijan	Bulgaria	Syria
3	-0.046	0.118	0.027
4	-0.364	0.022	-0.488
5	0.612	0.642	0.204
6	0.457	0.235	0.045
11	0.032	-0.002	-0.373
12	-0.007	-0.002	-0.298
21	0.187	0.281	-0.044
23	0.264	0.094	-0.411
26	0.302	-0.358	-0.373
27	-0.007	-0.640	-0.667
28	0.379	-0.640	-0.606
44	0.110	-0.150	0.134
51	-0.007	0.070	0.397
53	0.032	0.165	0.045
55	0.302	-0.252	0.151
57	-0.530	-0.100	0.045

In Table 3 items 3, 51 and 57 are in favour of Azerbaijani; items 44 and 55 are in favour of Bulgarian; items 5, 6, 11, 12, 21 and 23 are in favour of Syrian students. According to this, items 3, 51 and 57 are easier only for Azerbaijani students, items 5, 6, 11, 12, 21 and 23 are easier only for Syrian students and items 44, 51 and 55 are easier only for Bulgarian students. Another finding is that some items are in favour of two countries. For instance items 4 and 53 are in favour of both Azebaijani and Syrian students whereas items 26, 27 and 28 are in favour of Bulgarian and Syrian students.

When items showing DIF are analyzed in terms of their cognitive properties they require the following skills:

- Items 3,4,5 and 6 require the ability to reach a conclusion by constructing meaning between related parts
- Items 11 and 12 require the ability to find the part of a meaningful whole
- Items 21 nd 23 require the ability to predict the whole that the parts construct
- Items 26,27 an 28 require the ability to predict the parts from the whole by three dimensional-thinking
- Items 44, 51, 53, 55 and 57 require the ability to reach a conclusion by combining the given parts using the knowledge of arithmetic operation, letter and symbol according to a rule.

Similarly, Maller (2001) investigated the DIF of the Wechsler Intelligence Scale for Children–Third Edition (WISC-III). The WISC-III national standardization sample (N = 2200) was used to determine DIF in six WISC-III subtests. After fitting two parameter logistic and graded response models to the data, the items were tested for DIF using the DIF detection method based on item response theory likelihood ratio. Of the 151 items studied, 52 were found to function differently across the groups.

Carman and Taylor (2010) examined the relationship between the Naglieri Nonverbal Ability Test (NNAT), ethnicity and gender, as well as the socioeconomic status and NNAT performance. Correlations and multiple regression were used to examine the relationships between ethnicity, SES, and NNAT performance in a large kindergarten sample. The results suggest a significant relationship between ethnicity, SES, and NNAT performance.

As presented in Table 2, the instability statistics value in terms of gender was not significant. Therefore, it can be stated that TOBT did not contain DIF in terms of gender. Since there is no significant difference between females and males (*p > 0.001*), a Rasch tree cannot be produced. In other words, no partition occurs, which is shown in Figure 2.

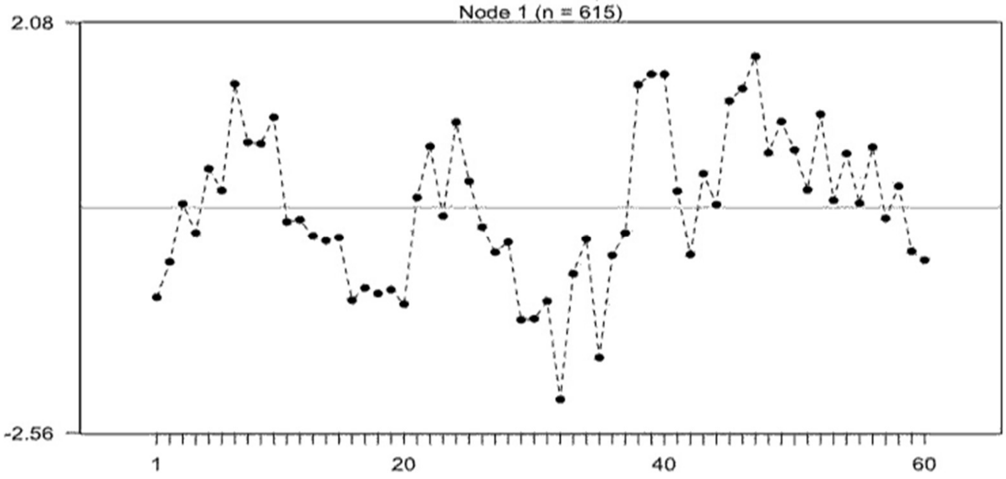


Figure 2. Rasch tree by gender

When Figure 2 is examined, it can be seen that there is no partition between females and males due to the absence of variation in response to the items in TOBT. This indicates that no item in TOBT shows DIF by gender.

Toivainen et al. (2017) used a large longitudinal twin sample to estimate sex differences in non-verbal and verbal abilities over time, using a variety of measures. Their study also investigated the influence of prenatal testosterone on these differences by comparing females with male co-twins to females with female co-twins. The sample size used in that study varied from 14187 participants at age 4 to 4959 participants at age 16. One-way ANCOVAs were used to establish significant group differences, either between sexes or between sex-by-zygosity twin groups. In all analyses, age was used as a covariate to account for the possible effect of age differences. The results showed negligible sex differences in non-verbal and verbal ability across development.

Empirical research consistently finds that standardized cognitive tests are not biased in terms of predictive and construct validity. Furthermore, continued claims of test bias, which appear in academic journals, the popular media, and some psychology textbooks, are not empirically justified. These claims of bias should be met with skepticism and evaluated critically according to established scientific principles (Brown, Reynolds, & Whitaker, 1999).

4. CONCLUSION

This study investigated whether recursively partitioned manifest variables can reveal DIF patterns in a non-verbal test using a Rasch tree approach. As a result of the research, 16 items in AYOS-TOBT 2017 indicated DIF in terms of countries. Concerning the 16 items showing DIF as a whole, it is noteworthy that the level of difficulty of the items according to the countries was close to the average value of zero. This situation showed that there was no significant variation in terms of countries.

In this study, whether the items in test showed DIF in terms of countries and gender was determined using quantitative analysis methods. Further research can be undertaken on the items which show DIF. Thus, comprehensive information can be obtained about the reasons why these items show DIF.

Acknowledgement

This article is the extended version of the paper titled "Investigating Differential Item Functioning of the Ankara University Examination for Foreign Students by Recursive Partitioning Analysis in the Rasch Model" and presented at the 6th International Congress on Measurement and Evaluation in Education and Psychology (C-MEEP) in 5 - 8 September 2018 in Prizren, Kosovo.

ORCID

Özge Altıntaş https://orcid.org/0000-0001-5779-855X
Ömer Kutlu https://orcid.org/0000-0003-4364-5629

5. REFERENCES

Altıntaş, Ö. (2016). *Ankara Üniversitesi Yabancı Uyruklu Öğrenci Seçme Testinin Ölçme Değişmezliğinin Örtük Sınıf ve Rasch Modeline Göre İncelenmesi [Investigating The Measurement Invariance of Ankara University Foreign Student Selection Test by Latent Class and Rasch Model]*. PhD diss., Ankara University.

AERA., APA., & NCME. (2014). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.

Anastasi, A. (1979). *Fields of Applied Psychology*. 2nd ed. New York: McGraw-Hill Book Company.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing*. 7th ed. New Jersey: Prentice-Hall International, Inc.

Ankara University. (2013). *Ankara Üniversitesine Yurtdışından Öğrenci Kabulüne İlişkin Yönerge. Ankara Üniversitesi Senato Kararı Örneği [Instruction on Admission of Foreign Students to Ankara University. Ankara University Senate Decision Example]*. Decision Date: 23.06.2013. Number of Meetings: 365. Number of Decisions: 3100.

ANKUDEM. (2011). *Ankara University Examination for Foreign Student Selection and Placement Exam (AYOS) Project Final Report*. Project No: 11Y5250001. Ankara: Ankara University Scientific Research Projects Scientific Coordinator.

Asimov, I. 2004. *Asimov's Chronology of Science and Discovery: Updated and Illustrated*. Norwalk: The Easton Press.

Best, J. W. (1970). *Research in Education*. 2nd ed. New Jersey: Prentice-Hall Inc.

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since Bias in Mental Testing. *School Psychology Quarterly*, *14*(3), 208-238. DOI: http://dx.doi.org/10.1037/h0089007

Büyüköztürk, Ş., Akgün, Ö. E., Demirel, F., Karadeniz, Ş., & Kılıç Çakmak, E. (2015). *Bilimsel Araştırma Yöntemleri (Scientific Research Methods)*. Ankara: Pegem Academy Publishing Co.

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. California: Sage Publication, Inc.

Carman, C. A., & Taylor, D. K. (2010). Socioeconomic Status Effects on Using the Naglieri Nonverbal Ability Test (NNAT) to Identify the Gifted/Talented. *Gifted Child Quarterly*, *54*(2), 75–84. DOI: https://doi.org/10.1177/0016986209355976

Cohen, R. J., Montague, P., Nathanson, L. S., & Swerdlik M. E. (1988). *Psychological Testing: An Introduction to Tests and Measurement*. California: Mayfield Publishing Company.

Crocker, L., & Algina, J. (1986). *Introduction to Classical Modern Test Theory*. New York: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. (1971). Test Validation. In *Educational Measurement*. 2nd ed., edited by R. L. Thorndike, 443-507. Washington: American Council on Education.

Cronbach, L. J. (1990). *Essentials of Psychological Testing*. 5th ed. New York: Harper and Collins Publishers, Inc.

Cronbach, L. J., & Meehl, P. E. (1955). Classics in the History of Psychology. *Psychological Bulletin*, *52*, 281-302. Retrieved from https://psychclassics.yorku.ca/Cronbach/construct.htm

Hambleton, R. K. (2002). Adapting Achievement Tests into Multiple Languages for International Assessments. In *Methodological Advances in Cross-National Surveys of Educational Achievement*, edited by A. C. Porter and A. Gamoran, 58-79. Washington: National Academy Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D., eds. (2005). *Adapting Educational and Psychological Tests for Cross-cultural Assessment.* New Jersey: Lawrence Erlbaum Associates, Inc.

Karakaya, İ., & Kutlu, Ö. (2012). An Investigation of Item Bias in Turkish Sub Tests in Level Determination Exam. *Education and Science, 37*(165), 2-15. Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/view/1342/433

Karasar, N. (2015). *Bilimsel Araştırma Yöntemi [Scientific Research Method]*. 28th ed. Ankara: Nobel Academy Publishing Co.

Kopf, J. (2013). *Model-based Recursive Partitioning Meets Item Response Theory: New Statistical Methods for the Detection of Differential Item Functioning and Appropriate Anchor Selection*. PhD diss., Munich: Ludwig-Maximilians-University Department of Statistics. Retrieved from https://edoc.ub.uni-muenchen.de/16434/1/Kopf_Julia.pdf

Kopf, J., Augustin, T., & Strobl, C. (2010). *The Potential of Model-Based Recursive Partitioning in the Social Sciences: Revisiting Ockham's Razor*. Technical report number 88. Munich: University of Munich Department of Statistics. Retrieved from https://pdfs.semanticscholar.org/c8ee/da16de9cb040066e5cb64aa37ceb58493286.pdf

Kutlu, Ö., & Karakaya, İ. (2007). Orta Öğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavının Faktör Yapılarına İlişkin Bir Araştırma. [A Research on the Factor Structure of Secondary Education Institutions' Student Selection and Placement Test]. *Elementary Education Online, 6*(3), 397-410. Retrieved from http://dergipark.gov.tr/download/article-file/90996

Linn, R. L., ed. (1989). *Educational Measurement*. 3rd ed. New Jersey: American Council on Education and Macmillan Publishing Company.

Linn, R. L., & Gronlund N. E. (2000). *Measurement and Assessment in Teaching*. 8th ed. New Jersey: Prentice-Hall International, Inc.

Lohman, D. F., & Lakin, J. M. (2011). Reasoning and Intelligence. In *Handbook of Intelligence*, edited by R. J. Sternberg, & S. B. Kaufman, 419-441. New York: Cambridge University Press.

Maller, S. J. (2001). Differential Item Functioning in the WISC-III: Item Parameters for Boys and Girls in the National Standardization Sample. *Educational and Psychological Measurement, 61*(5), 793–817. DOI: https://doi.org/10.1177/00131640121971527

Messick, S. (1989). Validity. In *Educational Measurement*. 3rd ed., edited by R. L. Linn, 13-103. New Jersey: American Council on Education and Macmillan Publishing Company.

Oral, T. (1985). *Lise Başarı Ölçüleri ile ÖSYS Puanları Arasındaki Uyum [Concordance between High School Success Metrics and University Entrance Exam (ÖSYS) Scores]*. PhD diss., Hacettepe University.

Osterlind, S. J., & Everson, H. T. (2009). *Differential Item Functioning*. California: Sage Publications, Inc.

Öner, N. (1987). Kültürlerarası Ölçek Uyarlamasında Bir Yöntembilim Modeli [A Methodology Model in Intercultural Scale Adaptation]. *Turkish Journal of Psychology*, *6*(21), 80-83.

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Resing, W. C. M. (2005). Intelligence Testing. In *Encyclopedia of Social Measurement*, edited by K. Kempf-Leonard, 307-315. San Diego: Elsevier Academic Press.

Reynolds, C. R., & Suzuki, L. (2013). Bias in Psychological Assessment: An Empirical Review and Recommendations. In *Handbook of Psychology Vol. 10: Assessment Psychology*. 2nd ed., edited by J. R. Graham, J. A. Naglieri, & I. B. Weiner, 82-113. New Jersey: John Wiley and Sons, Inc.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of Procedures for Detecting Test-Item Bias with Both Internal and External Ability Criteria. *Journal of Educational and Behavioral Statistics*, *6*(4), 317-375. DOI: https://doi.org/10.3102/10769986006004317

Sternberg, R. J. (1997). The Concept of Intelligence and Its Role in Lifelong Learning and Success. *American Psychologist*, *52*(10), 1030-1037. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.826.5234&rep=rep1&type=pdf

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch Trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289-316. Retrieved from https://eeecon.uibk.ac.at/~zeileis/papers/Strobl+Kopf+Zeileis-2015.pdf

Thornell, J. G., & McCoy, A. (1985). The Predictive Validity of The Graduate Record Examination for Subgroups of Students in Different Academic Disciplines. *Educational* *and Psychological Measurement*, *45*(2), 415-419. DOI: http://dx.doi.org/10.1177/001316448504500229

Toffler, A. (1980). *The Third Wave*. New York: McGraw-Hill Book Company.

Toker, F. (1997). *Türkiye'de Yükseköğretime Giriş [Entrance to Higher Education in Turkey]*. Ankara: ÖSYM Publications.

Toivainen, T., Papageorgiou, K. A., Tostoc, M. G., & Kovas Y. (2017). Sex Differences in Non-verbal and Verbal Abilities in Childhood and Adolescence. *Intelligence*, 64, 81-88. DOI: https://doi.org/10.1016/j.intell.2017.07.007

Turgut, M. F. & Baykul, Y. (1992-1). *Ölçekleme Teknikleri [Scaling Methods]*. Ankara: ÖSYM Publications.

USYM. (1978). *İki Aşamalı Üniversitelerarası Seçme ve Yerleştirme Sınavının Temel İlkeleri [Basic Principles of Two-Stage University Selection and Placement Exam]*. Ankara: ÖSYM Publications.

USYM. (1980a). *İki Aşamalı Üniversitelerarası Seçme ve Yerleştirme Sistemi [Two-Stage University Selection and Placement System]*. KY-02-80-0001. Ankara: ÖSYM Publications.

USYM. (1980b). *İki Aşamalı Üniversitelerarası Seçme ve Yerleştirme Sistemi [Two-Stage University Selection and Placement System]*. KY-02-80-0002. Ankara: ÖSYM Publications.

Walsh, W. B., & Betz, N. E. (1995). *Tests and Assessment*. 3rd ed. Englewood Cliffs. New Jersey: Prentice-Hall International, Inc.

Wicherts, J. M. (2007). *Group Differences in Intelligence Test Performance*. PhD diss., University of Amsterdam. Retrieved from https://pure.uva.nl/ws/files/4175964/46967_Wicherts.pdf

YOK. (2013). *Yurtdışından Öğrenci Kabulüne İlişkin Esaslar [Principles for the Acceptance of Students From Abroad]*. The decision of the General Council of Higher Education. Decision Date: 01.02.2013.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*. *17*(2), 492-514.

Zeileis, A., Strobl, C., Wickelmaier, F., & Kopf, J. (2011). *psychotree: Recursive partitioning based on psychometric models. R package version 0.12-1*, Retrieved from http://CRAN.R-project.org/package=psychotree

Zwick, R., & Sklar, J. C. (2005). Predicting College Grades and Degree Completion Using High School Grades and SAT Scores: The Role of Student Ethnicity and First Language. *American Educational Research Journal*, *42*, 439-464. DOI: https://doi.org/10.3102/00028312042003439

International Journal of Assessment Tools in Education

2019, Vol. 6, No. 4, 617–635

https://dx.doi.org/10.21449/ijate.593636

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

Development of a Preschool Teachers' Pedagogical Content Knowledge Scale regarding Mathematics

Hatice Dağlı [1,*], **H. Elif Dağlıoğlu** [1], **E. Hasan Atalmış** [1]

[1]Kahramanmaraş Sütçü İmam University, Department of Preschool Education, Turkey
[2]Gazi University, Department of Preschool Education, Turkey
[3]Kahramanmaraş Sütçü İmam University, Department of Measurement and Evaluation in Education, Turkey

ARTICLE HISTORY

Received: 18 July 2019
Revised: 07 October 2019
Accepted: 05 December 2019

KEYWORDS

Mathematics education,
Preschool teacher,
Pedagogical content knowledge

Abstract: This study aimed to develop a measurement tool in order to assess preschool teachers' pedagogical content knowledge regarding mathematics. The study was based on 300 preschool teachers working in formal independent kindergartens and nursery classes of primary/secondary schools in the Kahramanmaraş Province of Turkey. Among the participants, 150 were chosen for pre-application and 150 for the main application. The scale consists of five different case studies and a total of 35 items, including dialogues that focus on mathematical content and processes reflected in children's talk during their play. In calculating the reliability of the scale, Cronbach Alpha was found to be .95 for the pre-application and .96 for the main application. For the validity of the scale, exploratory and confirmatory factor analyses were performed. The exploratory factor analysis results revealed the scale to be a single-factor structure. When the factor loads of each relevant item were examined, no item was found to exist with a factor load value of less than .30. After confirmatory factor analysis was performed, the model fit indices of CFI, TLI, RMSEA, and SRMR values were found to be .91, .91, .06 and .06, respectively. These results show the model to be reliable to an acceptable level. Based on the findings, it could be concluded that the scale is an instrument that produces valid and reliable measures, and that it can be used in order to determine the preschool teachers' pedagogical content knowledge regarding mathematics.

1. INTRODUCTION

Mathematics and mathematical thinking have been regarded as key skills of our time in terms of their scientific field. The development of mathematical competencies begins at birth (Anthony & Walshow, 2009; Çoban, 2002). Mathematics is a field containing important concepts and skills which are widely used in learning processes and particularly in daily life. As people interact with their environment in daily life, they encounter various concepts such as time, space, shapes, and numbers, and therefore interact with mathematics without even realizing it (Bulut & Tarım, 2006). Understanding mathematics provides children with the ability to solve problems and to make correct decisions. Mathematics knowledge requires many

CONTACT: Hatice Dağlı ✉ hnddagli@hotmail.com Kahramanmaraş Sütçü İmam University, Department of Preschool Education, 46100, Kahramanmaraş, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

skills such as establishing cause-effect relationships, making calculations, calculating time, money management, and the use of technology (Ontario Ministry of Education [OME], 2014).

Researchers who conduct studies on cognitive development have revealed that the early development of mathematical skills closely relates to children's academic achievement in subsequent years (Anders & Rossbach, 2015; Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Clements, Sarama, & DiBiase, 2004; Gersten et al., 2009). Mathematics is a particularly hierarchical subject, in which mastery of simple concepts and procedures is required in order to understand more difficult mathematics (Watts, Duncan, Clements, & Sarama, 2017). It is of significant importance to teach mathematics that children will use and encounter throughout their lives (OME, 2014). In this regard, the early childhood years are particularly vital as the starting point for children to encounter formal mathematics education and basic mathematical concepts; moreover, mathematical skills are also learned within this period. The aim of mathematics education at the preschool level is to provide children with meaningful experiences through gameplay, stories, music and physical activities; to create them a sense of success with appropriate materials in appropriate physical environments, and to support the development of mathematics skills without creating a negative attitude towards mathematics (Arnas, 2006; Dağlı & Dağlıoğlu, 2017; Henniger, 1987; Metin, 1994; Mononen, Aunio, & Koponen, 2014). Thus, the preschool period is considered to the magical years where children's love for mathematics is inculcated and nurtured, and for the development of a positive attitude towards mathematics.

Some experimental studies on mathematical concepts and skills in the preschool period depicted that mathematical applications performed with children may create positive differences in their mathematical competencies when they start primary school, and that these differences last throughout their school life and even beyond (Anders, Grosse, Rossbach, Ebert, & Weinert, 2013; Sammons et al., 2004). In this context, mathematics literacy and mathematics skills are important not only for children's school success, but also in terms of their professional career throughout adulthood (Anders & Rossbach, 2015; Clements et al., 2004). Mathematics education presented to children during their preschool period is significant for their ability to achieve successful mathematical thinking in the following years and in their readiness preparation for primary schooling (Claesens & Engel, 2013; Dağlıoğlu, Dağlı, & Kılıç, 2013). When considered in the long term, understanding mathematics is so effective that it can direct children towards their future work life and career (Ontario Ministry of Education, 2014).

When it comes to the significance of early mathematics education, recent studies have emphasized that such education should be structured appropriately to the nature of the child; and that the child should reach the information by doing and experiencing personally, rather than teachers attempting to transfer knowledge directly to the child (Arnas, 2006; National Council of Teachers of Mathematics [NCTM], 2000). In other words, it is necessary for children to encounter the experiences in which they will learn mathematics concepts by doing and experiencing during their preschool period (Clements & Sarama, 2014; Umay, 2003).

The recent studies have also suggested that activities prepared in accordance with children's interests and their motivation can have a significant effect on their future success (Baranek, 1996; Berhenge, 2013; Mokrova, 2012; Tella, 2007). Education given in subjects or areas that children are interested in has a more lasting effect (Fisher, 2004). In this regard, mathematics content can and should play a significant role in early childhood education.

Different research on mathematics education during the preschool period has been conducted in many different countries. In the Turkish Preschool Education Program, which was updated in 2013, it is emphasized that mathematics education contributes to the cognitive development of children, that mathematics education within the preschool period can bestow positive attitudes in children, and that the mathematical inquiry skills of children can be improved through mathematics-based activities (Milli Eğitim Bakanlığı [Turkish Ministry of National Education], 2013). In addition, mathematical activities that establish relationships between concepts and life skills should be included in preschool education programs and that child-centered, game-based and multifaceted activities should be planned.

High quality, interesting and accessible mathematics education for the 3-6-year-olds age group was emphasized through situational assessments undertaken jointly by the National Association for the Education of Young Children in America and the NCTM (2010). In particular, the NCTM emphasized that educational programs which are well-planned, comprehensive, suitable for children's development, and meet the required language skills within a cultural context will be more effective (NCTM, 2009, 2013). Accordingly, mathematical understanding, knowledge and skills need to be gained during the education period starting from preschool. The NCTM (2009) also set content and process standards; defined the concepts and contents that children should learn through the content standards, and concept and content knowledge acquisition as well as using methods information through the process standards. When based on mathematics education, the NCTM (2000) showed that mathematical activities and mathematical content such as numbers, operations, geometry and measurement should be integrated with process standards such as problem solving, reasoning and proof, association, communication and symbolization. This process showed that mathematics program and education practices should be structured on a sound basis by taking into account both mathematical content areas and the developmental characteristics of children.

Considering that pedagogical approaches supporting the development of mathematical skills are seen as effective in enhancing these skills in children (Mononen & Aunio, 2013); the importance of teachers' pedagogical content knowledge related to mathematics has become prominent (Gifford, 2005). Pedagogical content knowledge in education was originally proposed by Shulman (1986), and encompasses knowing what to teach according to age groups and integrating that with the knowledge of how to teach it.

McCray (2008) explained the factors affecting pedagogical content knowledge regarding mathematics, as illustrated in Figure 1.

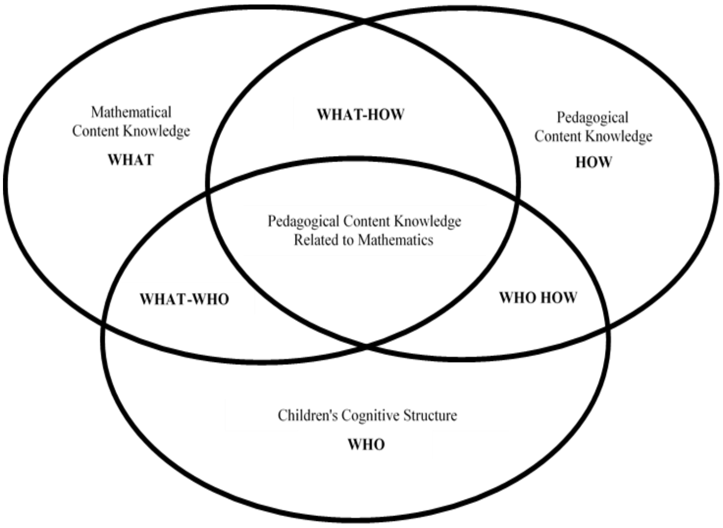


Figure 1. Pedagogical Content Knowledge Related to Mathematics (as revised by McCray, 2008)

McCray (2008) defined pedagogical content knowledge regarding mathematics as a junction point of three questions in mathematics education; Who will teach?, What to teach?, and How to teach? Teacher's pedagogical content knowledge, content knowledge and teaching ability are of paramount importance in effective learning and success in children (Jang, 2013; Zhang, 2015). The basis of mathematics education begins with an understanding of mathematical knowledge (Zhang, 2015).

Teachers with pedagogical mathematics concept knowledge know which concepts are the most basic and the best analogies that can gain help conceptual understanding; can enter events with new ideas in accordance with the interests of children, and ensure children use mathematics and mathematical language by asking children the right questions (McCray, 2008). The language used by teachers in the classroom should be founded upon improving the mathematical thinking of children. Guidelines for teachers in terms of mathematics contents such as numbers, spatial relationships, and operations should help children to use mathematics (Clements & Sarama, 2014; McGrath, 2010). Teachers should provide support to enable children to develop a positive attitude towards mathematics by taking full account of mathematics education and preparing appropriate programs in this regard (Copley, 2010; Dağlıoğlu, Genç, & Dağlı, 2017).

Previous studies show that teachers' attitudes, pedagogical field knowledge and beliefs affect children's mathematical ability, that the methods and techniques used by teachers affect children's ability in this field, and that teachers are lacking in mathematics education and in recognizing children's abilities (Chace Pierro, 2015; Cox, 2011; Erdoğan, 2006; Güven, 1998; Hacısalihoğlu Karadeniz, 2011; Kilday, 2010). From analyzing the relevant literature, a few studies have been specifically focused on preschool teachers' pedagogical content knowledge regarding mathematics (Cox, 2011; Kilday, 2010; McCray, 2008; Platas, 2008). However, studies that have been conducted in Turkey usually focus on preschool teachers' attitudes, beliefs and self-efficacy towards mathematics education (Çelik, 2017; Güven, Karataş, Öztürk, Arslan, & Gürsoy, 2013; Karakuş, Akman, & Ergene, 2018; Koç, Sak, & Kayri, 2015; Şeker & Alisinanoğlu, 2015); whereas only two studies considered pedagogical content knowledge regarding mathematics (Aksu & Kul, 2017; Parpucu & Erdoğan, 2017).

The current research was planned in order to develop a measurement tool for determining preschool teachers' pedagogical field knowledge of mathematics in order to address a gap in this field.

2. METHOD

This section includes information related to the working group, the data collection tool, and the development process of the scale.

2.1. Working Group

The participants of the study consisted of 300 teachers working in formal independent kindergartens and nursery classes of primary/secondary schools under the Turkish Ministry of National Education in Kahramanmaraş Province, Turkey; specifically, the districts of Dulkadiroğlu and Onikişubat. Of the participant teachers, 150 were selected for the pre-application and 150 for the main application.

While determining the size of the group to conduct factor analysis in the preliminary application, the researchers proposed different approaches; some argued that there should be twice the number of items (Büyüköztürk, Kılıç-Çakmak, Akgün, Karadeniz, & Demirel, 2008), some four times the number of items (MacCallum, Widaman, Preacher, & Hong, 2001), and others suggested 10 times (Nunnally, 1978). In addition, for exploratory factor analysis, Kaiser-Meyer-Olkin Test is expected to be greater than .50 and Bartlett test should be statistically significant (Büyüköztürk, 2010). In this regard, the decision was made to conduct an application with 150 teachers as the pre-application stage. Table 1 presents the demographic information regarding the participants.

Table 1. Participants' Demographic Information

Demographic Information			Pre-application		Main Application	
			Frequency	Percentage	Frequency	Percentage
Gender	Female		130	86.7	150	100.0
	Male		20	13.3	-	-
Age (years)	Less than 20		40	26.7	14	9.3
	21-25		47	31.3	42	28.0
	26-30		45	30.0	60	40.0
	31-35		18	12.0	33	22.0
	36 or over		-	-	1	0.7
Graduation	High School / Girls' Vocational High School		-	-	-	-
	Associate Degree	Child Development Vocational School	23	15.3	8	5.3
	Undergraduate	Preschool Teacher	81	54.0	99	66.0
	Undergraduate	Child Development	41	27.3	34	22.7
	Postgraduate		5	3.4	6	4.0
	Other		-	-	3	2.0
Seniority (years)	Less than 1 year		25	16.7	4	2.7
	1-5 years		36	24.0	23	15.3
	6-10 years		59	39.3	64	42.7
	11-15 years		18	12.0	37	24.6
	16-20 years		8	5.3	13	8.7
	21 years or more		4	2.7	9	6.0
Institution	Independent kindergarten		106	70.7	90	60.0
	Primary school		27	18.0	50	33.3
	Secondary school		17	11.3	10	6.7
Number of Children (per class)	5-10		2	1.3	5	3.3
	11-15		12	8.0	14	9.3
	16-20		74	49.3	67	44.7
	21 or more		62	41.3	64	42.7
Age of Children	36-53 months		15	10.0	-	-
	54-60 months		70	46.7	308	51.3
	61-66 months		65	43.3	292	48.7
Mathematical Activities	Never		-	-	-	-
	One time per 2-3 weeks		-	-	3	2.0
	Twice a week		41	27.3	38	25.3
	Three to four times a week		86	57.3	73	48.7
	Daily		23	15.4	36	24.0

Table 1 shows that 86.7% (*n* = 130) of the participant teachers were female, whilst 13.3% (*n* = 20) were male for the pre-application stage; whereas, all participants were female for the main application. In the pre-application, 10% (*n* = 15) of the teachers were working with children aged between 36 and 53 months, 46.7% (*n* = 70) between 54 and 60 months, and 43.3% (*n* = 65) between 61 and 66 months. In the main application, 51.3% (*n* = 308) of the teachers were working with children aged between 54 and 60 months, whilst the other 49.7% (*n* = 292) were working with children aged between 61 and 66 months.

2.2. Data Collection Tool

The research data was collected through a "Teacher Information Form" and a developed "Preschool Teachers' Pedagogical Content Knowledge Scale regarding Mathematics

(PTPCKSM)" that included five case studies. The Teacher Information Form was used in order to record the teachers' gender, age, type of graduation school, their seniority, type of institution where they were assigned, the number of children in each class, the age group of the children in their class, and the availability of a mathematics center in class.

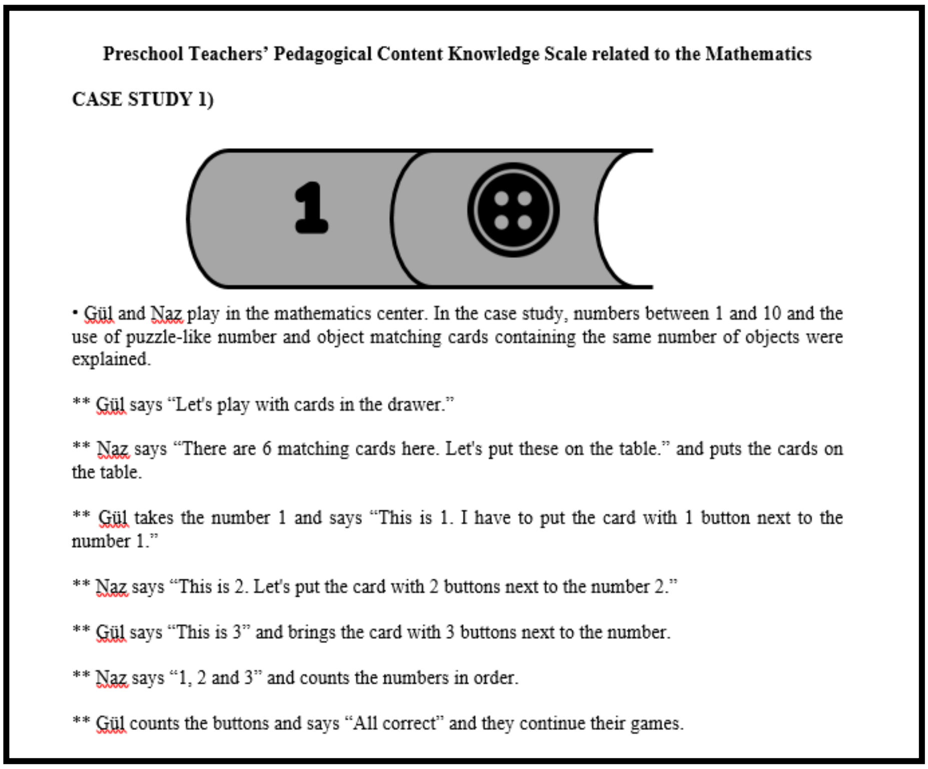
The PTPCKSM was developed in order to identify teachers' awareness towards mathematical content and the processes involved in language used by children. In this section, five case studies were designed based on children's dialogues including different mathematical contexts and processes from the expressions used by children during play. A separate marking form was created for each case study and teachers were requested to mark the mathematical contents and processes they identified in accordance with the form. Based on the NCTM (2000) standards, the case studies of the PTPCKSM included "counting, geometry, spatial perception, part–whole relationships, matching, classification/grouping, comparison, sorting, measurement, operation, pattern, and graphics" as the mathematics contents, and "communication, association, reasoning and proof, problem solving and representation/symbolization" as the mathematical processes. Each case study consisted of seven statements/items.

During the scale's development process, first the existing literature was reviewed. The contents and processes involved in mathematics education during the preschool period in Turkey and elsewhere were examined, and the sub-dimensions for the PTPCKSM were formed after determining the problem statement based on the aforementioned content. The scale known as "Knowledge of Mathematical Development" that was developed by Platas (2008) for the purpose of measuring teachers' knowledge on the development of mathematical concepts in children was taken as the basis for the current study. Along with the necessary permissions in the ongoing process, the "Preschool Mathematics-Pedagogical Concept Information Interview Form" that was developed by McCray and Chen (2008) was also taken into consideration.

Two of the five case studies in the PTPCKSM were prepared based on the Preschool Mathematics-Pedagogical Concept Information Interview Form; with the other three case studies formed by the researcher. The case studies were designed based on a straight line approach, from simple to complex. Each case study contained different yet simple images in order to add clarity to the case studies. With examination of the content and process standards developed by the National Association for the Education of Young Children in America and NCTM, as well as the involvement in the development process of mathematical concepts in children, significant attention was paid to the inclusion of these standards in case study. Within the scale development process, three different scale drafts were prepared in the form and coding dimension, and each draft was applied to three different preschool teachers. The scale was then finalized by testing the clarity of the scale with various applications.

The expert opinion of seven specialists in mathematics and preschool education, who were also faculty members at different universities, were obtained in the preparation of the scale. In addition, a measurement and evaluation expert plus and two Turkish linguists were employed to examine the scale in terms of the language clarity and application of the items. The scale was considered ready for the pre-application stage after having taken a total of nine expert opinions. In the PTPCKSM, spelling errors and incoherencies were corrected so as to increase the scale's clarity along with the expert opinion. In order that the data could be grouped and the correct comparisons made, International Standard Classification of Education (ISCED) categories (Türkiye İstatistik Kurumu Başkanlığı [Turkish Statistical Institute], 2012) were employed. In addition, categories (as mathematics contents and mathematical processes) were created for some related items.

One of the case studies is presented in Figure 2, and the scoring table in the marking form created for the teachers to record their answers is shown in Figure 3. In each case study, first the image and the text were taken as a whole; then, in the marking section, each sentence (item) was taken separately by dividing each sentence (item) in the case study. Here, the teachers were expected to see the whole, then the application was made by dividing the text into sentences in order to be more easily recognize the details in the text.



Preschool Teachers' Pedagogical Content Knowledge Scale related to the Mathematics

CASE STUDY 1)

• Gül and Naz play in the mathematics center. In the case study, numbers between 1 and 10 and the use of puzzle-like number and object matching cards containing the same number of objects were explained.

** Gül says "Let's play with cards in the drawer."

** Naz says "There are 6 matching cards here. Let's put these on the table." and puts the cards on the table.

** Gül takes the number 1 and says "This is 1. I have to put the card with 1 button next to the number 1."

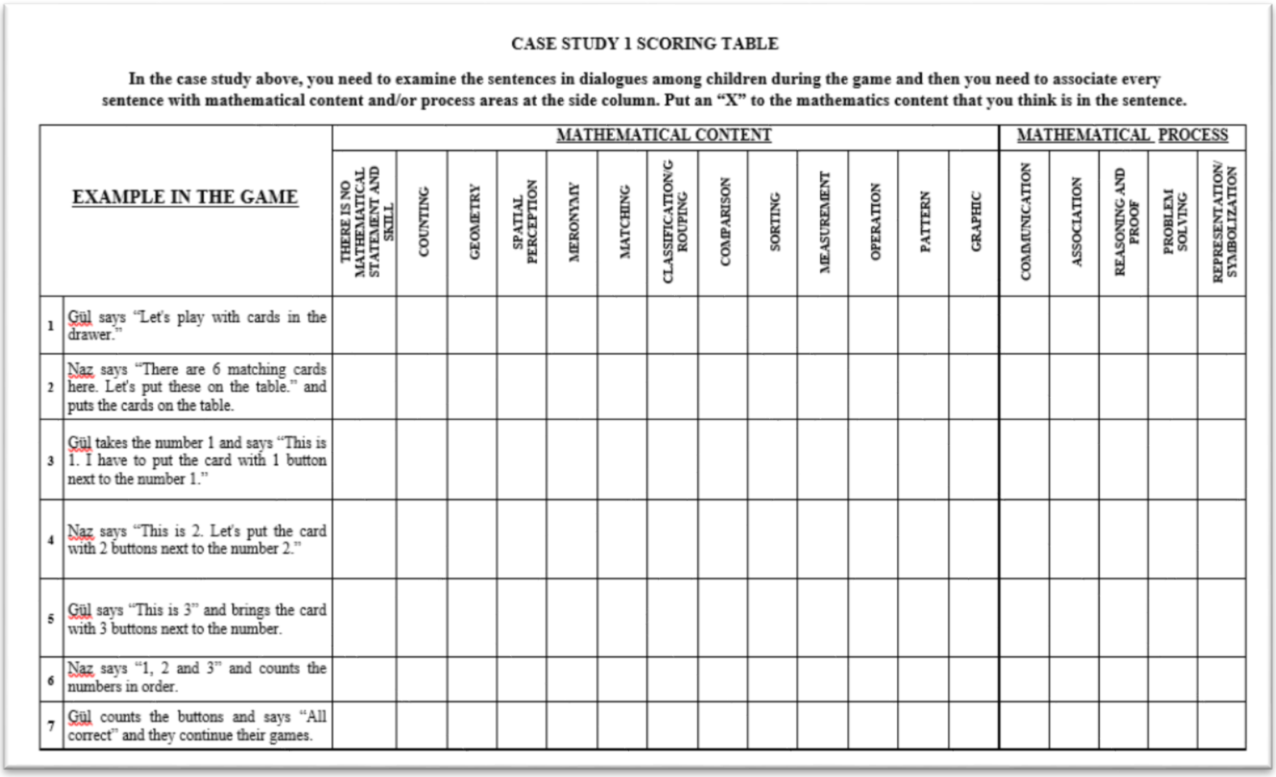
** Naz says "This is 2. Let's put the card with 2 buttons next to the number 2."

** Gül says "This is 3" and brings the card with 3 buttons next to the number.

** Naz says "1, 2 and 3" and counts the numbers in order.

** Gül counts the buttons and says "All correct" and they continue their games.

Figure 2. PTPCKSM Case Study Form



CASE STUDY 1 SCORING TABLE

In the case study above, you need to examine the sentences in dialogues among children during the game and then you need to associate every sentence with mathematical content and/or process areas at the side column. Put an "X" to the mathematics content that you think is in the sentence.

	EXAMPLE IN THE GAME	MATHEMATICAL CONTENT													MATHEMATICAL PROCESS				
		THERE IS NO MATHEMATICAL STATEMENT AND SKILL	COUNTING	GEOMETRY	SPATIAL PERCEPTION	MERONYMY	MATCHING	CLASSIFICATION/G ROUPING	COMPARISON	SORTING	MEASUREMENT	OPERATION	PATTERN	GRAPHIC	COMMUNICATION	ASSOCIATION	REASONING AND PROOF	PROBLEM SOLVING	REPRESENTATION/ SYMBOLIZATION
1	Gül says "Let's play with cards in the drawer."																		
2	Naz says "There are 6 matching cards here. Let's put these on the table." and puts the cards on the table.																		
3	Gül takes the number 1 and says "This is 1. I have to put the card with 1 button next to the number 1."																		
4	Naz says "This is 2. Let's put the card with 2 buttons next to the number 2."																		
5	Gül says "This is 3" and brings the card with 3 buttons next to the number.																		
6	Naz says "1, 2 and 3" and counts the numbers in order.																		
7	Gül counts the buttons and says "All correct" and they continue their games.																		

Figure 3. PTPCKSM Scoring Table

As can be seen in the case study shown in Figure 2 and Figure 3, the PTPCKSM contains one or more mathematical contents/processes within each sentence. Each item was rated as 1 point in the scale; therefore, the whole case study was calculated as a total of 7 points, with each case study consisting of seven items. The pedagogical content knowledge of the teachers with the highest total score were therefore expected to be considered as high.

2.3. Data Collection Process

In both the pre-application and the main application, educational institutions were visited by the researcher where the necessary permissions had been received. The school principals were informed about the study first, and then interviews were subsequently held with each participant teacher. The scale was introduced to the teachers in person by the researcher, and any necessary corrections to the scale were applied together. The teachers requested additional time in order to better complete the scale. The forms were delivered to the teachers by the researcher, and later retrieved according to pre-specified dates.

2.4. Data Analysis

IBM's SPSS 22 statistical package and the Mplus 7.4 program were used in order to calculate the reliability and validity of the developed scale. Cronbach Alpha coefficient was used to test the reliability of the scale, and then item difficulty index values and item discrimination coefficients were calculated separately for each item. In order to calculate the validity of the scale, the content and construct validity were examined; both exploratory and confirmatory factor analyses were performed for this purpose. Kaiser-Meyer-Olkin Test and Bartlett Tests were performed in exploratory factor analysis (EFA); whilst CFI, TLI, RMSEA and SRMR values were calculated for the scale's confirmatory factor analysis (CFA).

Before scoring, each mathematics content and process in the scale was alphabetically coded as a, b, c, d, e, f…….p. Table 2 shows the codes corresponding to the mathematical contents and processes.

Table 2. PTPCKSM Content/Process Coding Values

Mathematics Content/Process	Coding Value
There is No Mathematical Statement and Skill	a
Counting	b
Geometry	c
Spatial Perception	d
Part–Whole Relationships	e
Matching	f
Classification/Grouping	g
Comparison	h
Measurement	i
Operation	j
Pattern	k
Graphic	l
Communication	m
Association	n
Reasoning and Proof	o
Problem Solving	ö
Representation/Symbolization	p

In the scoring method, each item was awarded equal points (equal scoring) (Frary, 1989; Masters, 1988).

Each item is worth 1 point, and in scoring the item total is divided by the number of answers required for its content/process. For example, the per code value of Item 1, in which the correct answer was "d and m," is calculated as 1/2 (0.50); and the per code value of Item 2, in which the correct answer was "b, d and m," is calculated as 1/3 (0.33).

Any incorrect answer results in 1 point deducted from the total score. For example, a response of "b, d and n" for Item 3 includes one correct and two incorrect answers; therefore, the score corresponds to 2 - 1 correct answer is calculated as 1/3 (0.33) points because there are three correct answers in this question.

3. RESULTS

The research findings related to the preschool teachers' pedagogical content knowledge on mathematics are reported in the following figures and tables.

3.1. Results for Pre-Application

3.1.1. *Reliability*

Each case study in the PTPCKSM and the reliability coefficient calculation for the whole scale are shown in Table 3. Table 3 shows that the reliability coefficient for each case study was found to be more than .70, and that the reliability coefficient levels of the whole scale and each case study were therefore considered "high" (Büyüköztürk, 2010). The reliability coefficients of the case studies were identified as varying from .94 to .96; and the reliability coefficient for the whole scale was found to be .95.

Table 3. Reliability Coefficients of PTPCKSM

Case Study	Number of Items	Reliability coefficient
Case Study 1	7	.95
Case Study 2	7	.94
Case Study 3	7	.96
Case Study 4	7	.94
Case Study 5	7	.96
Whole Scale	35	.95

Item difficulty and discrimination indices for the items in each case study are presented as shown in Table 4.

Table 4. Item-level Statistics Related to PTPCKSM Case Studies

Item	Case Study 1		Case Study 2		Case Study 3		Case Study 4		Case Study 5	
	p	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*
Item 1	.49	.69	.70	.57	.56	.79	.53	.87	.51	.90
Item 2	.51	.85	.45	.91	.38	.93	.50	.71	.38	.90
Item 3	.42	.95	.37	.92	.35	.94	.36	.44	.43	.92
Item 4	.38	.88	.46	.84	.40	.92	.38	.91	.57	.81
Item 5	.40	.93	.38	.86	.39	.91	.45	.93	.45	.90
Item 6	.53	.78	.39	.90	.47	.85	.44	.91	.34	.72
Item 7	.42	.89	.36	.91	.37	.78	.47	.87	.45	.88

Table 4 shows the item difficulty index values (*p*) and the item discrimination index values (*r*) of the items in Case Study 1. Basol (2015) classified item difficulty as "extremely easy" (*p* = .85 to 1.00), "easy" (*p* = .61 to .84), "medium" (*p* = .40 to .60), "difficult" (*p* = .16 to .39), and "extremely difficult" (*p* = .00 to .15). For Case Study 1, the item difficulties differed from .38 to .53 and were therefore classed as either medium (*p* = .49, .51, .42, .40, .53, .42) or difficult (*p* = .38). The item discrimination index values in Case Study 1 varied between .69 and .95.

The item difficulty index values of Case Study 2, varied between .36 and .70, with item difficulties easy (*p* = .70), medium (*p* = .47, .40) or difficult (*p* = .39, .38, .37, .35). The item discrimination index values for Case Study 2 varied between .57 and .92.

The item difficulty index values of Case Study 3 varied between .35 and .56, with item difficulties either medium (*p* = .56, .46, .45) or difficult (*p* = .39, .38, .37, .36). The item discrimination index values for Case Study 3 varied between .78 and .94.

The item difficulty index values of Case Study 4 varied between .36 and .53, with item difficulties either medium (*p* = .53, .50, .47, .45, .44) or difficult (*p* = .38, .36). The item discrimination index values for Case Study 4 varied between .44 and .93.

The item difficulty index values of Case Study 5 varied between .34 and .57, with item difficulties either medium (*p* = .53, .51, .45, .43) or difficult (*p* = .38, .34). The item discrimination index values for Case Study 5 varied between .72 and .90.

The results indicated that the questions were classified as either difficult, medium or easy, and that the item discrimination index values were found to have more than .30 of variance explained by the scale (Thorndike, 2005).

3.1.2. *Validity*

The content and the construct validity indices were examined through exploratory and confirmatory factor analyses.

3.1.2.1. Content Validity

The opinion of seven experts was sought in order to assess the content validity of the PTPCKSM. All of the items were accepted by the experts.

The content validity rate for each item was determined based on the evaluation of the expert opinion. Afterwards, the content validity index value was determined by taking the average of the calculated rates. The index value for each item was then used by the experts to determine whether or not the item was deemed necessary (Büyüköztürk, 2010; Yurdugül, 2005).

The content validity index value was calculated for the eligibility level of the scale items as a whole. With seven experts, scales with a content validity index value of more than .99 can assure scope validity (Yurdugül, 2005). From calculation of the content validity index values for the PTPCKSM, the eligibility level of the items in terms of their intended purpose and the level of the children was calculated as "+1." This value shows that all items in the PTPCKSM were deemed to be necessary, and that the scale's content validity was assured as a whole.

3.1.2.2. Construct Validity

Exploratory factor analysis (EFA) of the scale was conducted in order to demonstrate the construct validity of the PTPCKSM at the pre-application stage. Both the Kaiser-Meyer-Olkin and Bartlett tests were performed to understand whether or not the scale was appropriate for factor analysis. For ensuring factor analysis of a scale, the Kaiser-Meyer-Olkin result should be .50 or above, and the Bartlett Sphericity result should be statistically significant ($p < .01$) (Büyüköztürk, 2010).

The analysis results showed that the Kaiser-Meyer-Olkin result for the PTPCKSM was .97 and that the Bartlett sphericity test (*p* < .01) was statistically significant. This result shows that factor analysis may be performed on the scale. Upon examining the eigenvalue for both methods, there are two factors that score as more than 1; with the first factor being 25.58 and the second factor 1.86. These two factors were found to account for 78.41% of the total variance in the scale, with 73.11% explained by Factor 1 and 5.30% by Factor 2. Considering the eigenvalue and the explained variance, Factor 1 was found to be about 14 times more dominant than Factor 2. This result therefore signified that the scale has a single factor structure. Figure 4 presents a scatter plot graph of the scale.

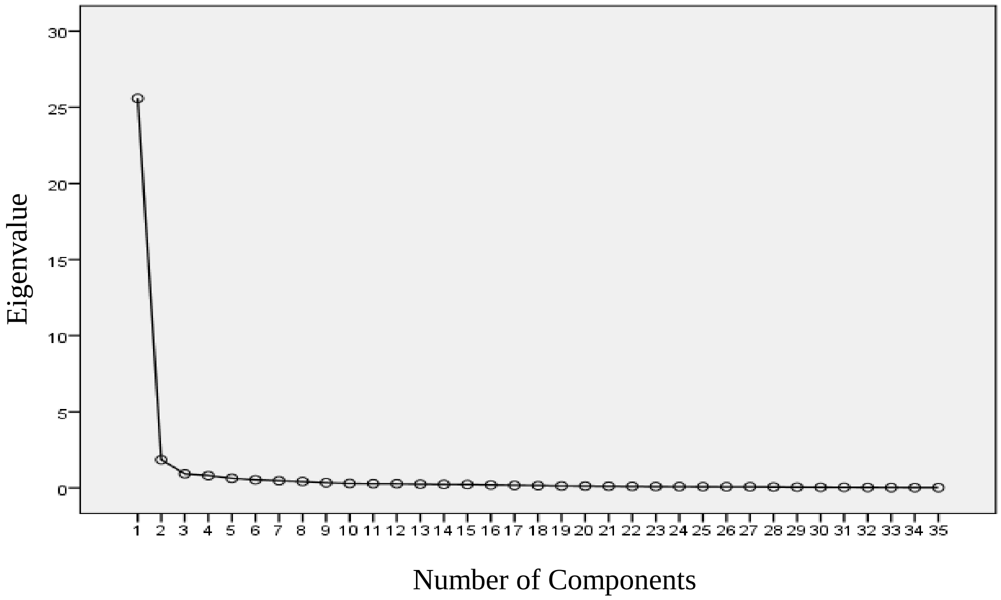


Figure 4. Scatter Plot Graph

Table 5. Factor Load Values as a Result of Principal Component Analysis of PTPCKSM

Item	Factor Loads		Item	Factor Loads	
	Factor 1	Factor 2		Factor 1	Factor 2
O1_3	.95		O1_7	.89	
O3_3	.94		O5_7	.89	
O4_5	.93		O1_4	.88	
O3_2	.93		O4_7	.87	
O1_5	.93		O4_1	.87	
O2_3	.93		O2_5	.87	
O3_4	.93		O3_6	.85	
O5_3	.92		O1_2	.85	
O3_5	.92		O2_4	.84	
O4_4	.91		O5_4	.81	
O2_2	.91		O1_6	.78	
O2_7	.91		O3_1	.78	
O5_2	.91		O3_7	.77	
O4_6	.91		O5_6	.72	
O5_1	.90		O4_2	.69	.62
O2_6	.90		O1_1	.69	
O5_5	.90		O2_1	.55	
			O4_3	.42	.81

When the factor loadings of each item were examined in the next stage, no item was found with a factor load value of less than .30, and therefore no items were removed from the scale. According to Table 5, only two factors were identified as being linked at the same time, such as Case Study 4, Item 2 and Case Study 4, Item 3. As the difference between the Factor 1 loading (.42) and Factor 2 loading (.81) of Item O4_3 was greater than .10, it was assumed that this problem was only due to Factor 2. As the difference between the Factor 1 loading (.69) and Factor 2 loading (.62) of Item O4_2 (Case Study 4, Item 2) was less than .10, it was considered that this item should be removed from the scale. However, since this item was thought to contribute to the scale contextually (content validity), it was decided not to remove the item from the scale.

3.2. Findings for the Main Application

As in the pre-application, the reliability coefficient was examined and the construct validity was also tested in the main application. However, the construct validity was tested by confirmatory factor analysis (CFA) in the main application.

3.2.1. *Reliability*

First, as shown in Table 6, the reliability coefficients of the scale were calculated at both the individual case study level and for the whole scale.

Table 6. Reliability Coefficients of PTPCKSM

Case Study	Number of Items	Reliability coefficient
Case Study 1	7	.91
Case Study 2	7	.73
Case Study 3	7	.82
Case Study 4	7	.81
Case Study 5	7	.86
Whole Scale	35	.96

According to the findings presented in Table 6, the reliability coefficient for each of the individual case studies was found to be more than .70, and that the reliability coefficient levels of the scale as a whole and each case study were considered to be high (Büyüköztürk, 2010). The reliability coefficients of the individual case studies varied from .73 to .91; and the reliability coefficient of the whole scale was found to be .96.

3.2.2. *Validity*

In order to test the construct validity of the scale at the next stage, confirmatory factor analysis (CFA) was performed using the MPlus 7.4 program, and the model produced by this analysis is shown as Figure 5. When the goodness of fit indices of the model, it was found that the CFI and TLI values were greater than .90, and that the RMSEA and SRMR values were less than .08. These results showed that the model was at an acceptable level (Kline, 2016). The χ^2/SD value was calculated to be less than the accepted value of 4 (χ^2 (547,150) = 821.76; CFI = .91; TLI = .91; RMSEA = .06; SRMR = .06). These results support that the scale has acceptable construct validity (Kline, 2016). As a result, both the reliability and validity analyses results for the main application revealed the PTPCKSM to be a suitable measurement tool.

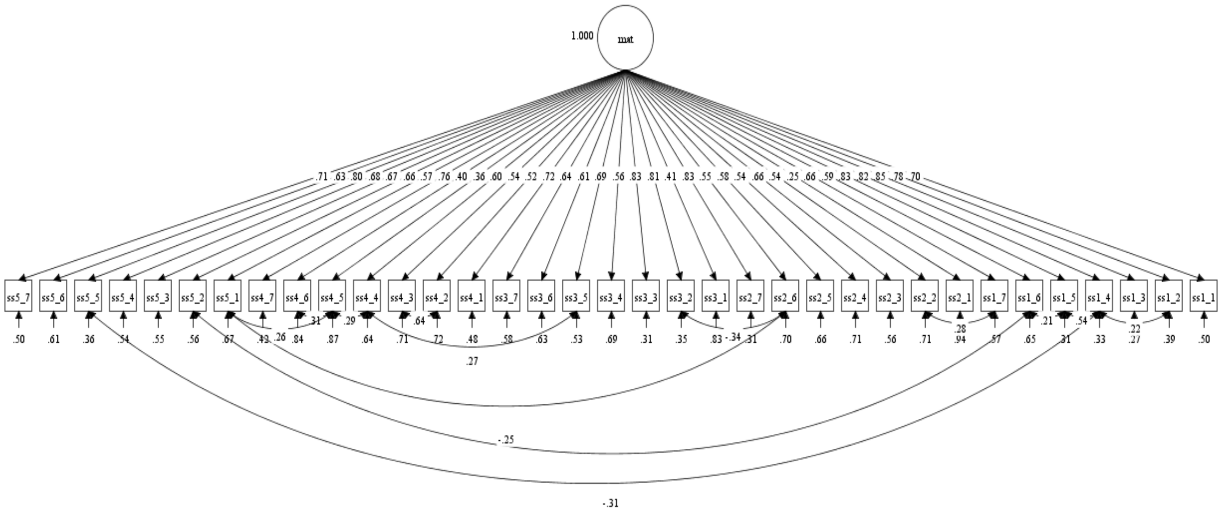


Figure 5. PTPCKSM Confirmatory Factor Analysis Model

4. DISCUSSION and CONCLUSION

Considering that mathematics is intertwined in our life skills and that the skills and processes related to mathematics develop in children during their early years, the importance of preschool teachers' pedagogical knowledge about mathematics is significant. This position illustrates the necessity for different measurement tools to assess preschool teachers' pedagogical content knowledge. Therefore, the current study was conducted in order to develop a new tool known as the Preschool Teachers' Pedagogical Content Knowledge Regarding Mathematics Scale; as well as to perform validity and reliability studies on the developed scale.

The participants of the study were 300 preschool teachers working in formal independent kindergartens and in nursery classes of primary/secondary schools under the Turkish Ministry of National Education within the Kahramanmaraş Province of Turkey.

A Teacher Information Form and the Preschool Teachers' Pedagogical Content Knowledge Regarding Mathematics Scale, which were both developed by the researcher, were employed as the data collection tools in this study. A pre-application study was conducted in order to determine the clarity and responsiveness of the PTPCKSM scale items, and all of the items were identified to have the necessary level of clarity.

For a valid scale, the problem should be well-defined, and statistically accepted values for both validity and reliability should be assured during preparation of the scale items (Büyüköztürk, 2005). The reliability coefficients of the PTPCKSM were calculated. The findings presented in Table 1 reveal that the reliability coefficient of each case study in the PTPCKSM to be more than .70, and that the reliability coefficient of the whole scale was .95 which indicated the scale to be reliable (Büyüköztürk, 2010). The item difficulty index value (*p*) and the item discrimination index value (*r*) were calculated separately for each item in each case study of the scale (see Table 2). When the item difficulty index values of the case studies were examined, they were found to vary between easy, medium and difficult; whilst the item discrimination index values were found to be higher than .30.

In addition to reliability, another requirement for determination of the scale is validity (Karasar, 2012). Therefore, explanatory and confirmatory factor analyzes indicating the construct validity as well as the content validity analysis were performed.

Seven expert opinions were consulted to determine the content validity of PTPCKSM. All of the items were welcomed by the experts. Content validity index was examined with the expert opinions; as a result of the calculation of the content validity index values of PTPCKSM, content validity index for eligibility level of the items in terms of the purpose and level of children was calculated as "+1". This value showed that all the items in the scale were necessary and the scale guaranteed content validity as a whole.

Kaiser-Meyer-Olkin Test result for PTPCKSM was found to be .97 and the Bartlett sphericity test (p <0.01) was statistically significant. This result showed that factor analysis can be performed on the scale. As a result of the factor analysis, Case Study 4 Item 2 and Case Study 4 Item 3 including contents and processes such as counting, geometry, classification/grouping and communication were linked two factors at the same time; as the difference between the two factors of the Case Study 4 Item 3 was greater than 0.10, and that this problem may be only due to factor 2. The difference between the two factors of the Case Study 4 Item 2 was found less than 0.10. However, since this item was assumed to contribute to the scale contextually (content validity) and the removal of the item would harm the integrity of the scale; this item was not removed from the scale.

In the next stage, confirmatory factor analysis was performed through use of the MPlus 7.4 program to test the construct validity of the scale (Figure 3). When model fit indexes were examined, CFI and TLI values were identified to be more than 0.90 and 0.90 and RMSEA and SRMR values were less than 0.08 and the model was in an acceptable level in terms of its construct validity (Kline, 2016). It was also found that the χ^2/SD value was less than the accepted value of 4 (χ^2 (487.152) = 972.22; CFI = .94; TLI = .93; RMSEA = .07; SRMR = .04). According to these applications, it was revealed that the Preschool Teachers' Pedagogical Content Knowledge Related to the Mathematics Scale is a suitable measuring tool including five case studies and 35 items based on the assessment of preschool mathematical contents and processes.

Mathematics is a hierarchical field, and basic mathematical skills and concepts are acquired during the preschool period. Therefore, children's acquirements within this scope come to the fore especially during their early childhood. The relevant literature revealed that the attitudes, approaches, beliefs and pedagogical content knowledge of preschool teachers related to mathematics are primary factors affecting children's acquirements in this field (Chace Pierro, 2015; Cox, 2011; Erdoğan, 2006; Güven, 1998; Hacısalihoğlu Karadeniz, 2011; Kilday 2010). However, considering the academic studies carried out with regards to preschool teachers' pedagogical content knowledge related to mathematics, only a few studies have been conducted on this subject in Turkey (Aksu & Kul, 2017; Parpucu & Erdoğan, 2017).

When studies on the pedagogical content knowledge related to mathematics for teachers working in primary education or upper education levels were examined, the knowledge level of teachers were shown to be low, and that major changes can be seen in teachers' thoughts and beliefs on mathematics education and teaching after having received supportive training in this area (Even & Tirosh, 1995; Gökkurt & Soylu, 2016; Nicol & Crespo, 2006; Tanışlı, 2013). Under these circumstances, the scale developed in the current study is expected to contribute to the related field. Based upon the study's findings, it is recommended that the validity and reliability of the developed scale be repeated for teachers working in different regions and provinces across Turkey. The scale may also be applied to teachers working with different age groups, in areas differentiated by socioeconomic level, and in different preschool education institutions. Relations between teachers' pedagogical content knowledge regarding mathematics and variables such as children's mathematical ability, levels of children's love for mathematics, and their attitudes towards mathematics may be analyzed and in-depth examinations conducted in order to reveal how teachers' content knowledge related to mathematics affects the development of children's mathematical concepts.

Acknowledgement

This paper is part of a doctoral dissertation of the corresponding author.

ORCID

Hatice Dağlı https://orcid.org/0000-0002-0788-0413
H. Elif Dağlıoğlu https://orcid.org/0000-0002-7420-815X
E. Hasan Atalmış https://orcid.org/0000-0001-9610-491X

5. REFERENCES

Aksu, Z., & Kul, U. (2017). Turkish adaptation of the survey of pedagogical content knowledge in early childhood mathematics education. *International Journal of Eurasia Social Sciences*, *8*(30), 1832-1848.

Anders, Y., Grosse, C., Rossbach, H., Ebert, S., & Weinert, S. (2013). Preschool and primary school influences on the development of children's early numeracy skills between the ages of 3 and 7 years in Germany. *School Effectiveness and School Improvement, 24*(2), 195-211.

Anders, Y., & Rossbach, H.G. (2015). Preschool teachers' sensitivity to mathematics in children's play: The influence of math-related school experiences, emotional attitudes, and pedagogical beliefs. *Journal of Research in Childhood Early Education, 29*(3), 305-322.

Anthony, G., & Walshaw, M. (2009). Mathematics education in the early years: Building bridges. *Contemporary Issues in Early Childhood, 10*(2), 107-121. doi: http://dx.doi.org/10.2304/ciec.2009.10.2.107

Arnas, Y. (2006). *Okul öncesi dönemde matematik eğitimi* [Mathematics education in preschool], Adana: Nobel.

Aunola, K., Leskinen, E., Lerkkanen, M.K., & Nurmi, J.E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology, 96*(4), 699-713.

Baranek, L. K. (1996). *The effect of rewards and motivation on student achievement.* Master's thesis. Grand Valley State University, MI.

Basol, G. (2015). *Eğitimde ölçme ve değerlendirme* [Measurement and evaluation in education], Ankara: Pegem.

Berhenge, A. L. (2013). *Motivation, self-regulation, and learning in preschool*. (Doctoral Dissertation). University of Michigan, USA.

Bulut, S., & Tarım, K. (2006). Okul öncesi öğretmenlerinin matematik ve matematik öğretimine ilişkin algı ve tutumları [*Perceptions and attitudes of preschool teachers about mathematics and mathematics teaching*]. *Çukurova Üniversitesi, Eğitim Fakültesi Dergisi*, *2*, 32-65.

Büyüköztürk, Ş. (2005). Anket geliştirme [Survey Development]. *Türk Eğitim Bilimleri Dergisi, 3*(2), 133-151.

Büyüköztürk, Ş. (2010). *Sosyal bilimler için veri analizi el kitabı. istatistik, araştırma deseni SPSS uygulamaları ve yorum. değerlendirme* [Data analysis handbook for social sciences statistics, research design SPSS applications and interpretation], Ankara: Pegem Akademi.

Büyüköztürk, Ş., Kılıç-Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2008). *Bilimsel Araştırma Yöntemleri* (1st Ed.) [Scientific Research Methods], Ankara: Pegem.

Chace Pierro, R. (2015). *Teachers' knowledge, beliefs, self-efficacy, and implementation of early childhood learning standards in science and math in prekindergarten and kindergarten.* (Master's thesis). University of North Carolina, USA. Retrieved from https://libres.uncg.edu/ir/uncg/f/Pierro_uncg_0154M_11772.pdf

Claesens, A., & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. *Teachers College Record, 115*(6), 1-29.

Clements, D. H., & Sarama, J. (2014). *Learning and teaching early math: The Learning Trajectories Approach* (2nd Ed.). New York, NY: Routledge.

Clements, D. H., Sarama, J., & DiBiase, A.-M. (Eds.). (2004). *Engaging young children in mathematics: Standards for early childhood mathematics education*. Mahwah, NJ: Erlbaum.

Copley, J. V. (2010). *The young child and mathematics* (2nd Ed.). Washington DC: National Association for the Education of Young Children.

Cox, G. J. (2011). *Preschool caregivers' mathematical anxiety: examining the relationships between mathematical anxiety, and knowledge and beliefs about mathematics for young children* (Doctoral Dissertation). Texas Woman's University, Denton, TX.

Çelik, M. (2017). Okul öncesi öğretmenlerin erken matematik eğitimine ilişkin öz yeterliklerinin çeşitli değişkenler açısından incelenmesi [Pre-school teachers' self efficacy related to early maths education]. *e-Kafkas Eğitim Araştırmaları Dergisi, 4*(1), 1-10.

Çoban, A. (2002). *Matematik dersinin ilköğretim programları ve liselere giriş sınavları açısından değerlendirilmesi* [Evaluation of mathematics course in terms of primary education programs and high school entrance exams]. In Proceedings of the V. National Science and Mathematics Education Congress, Middle East Technical University, Ankara, Turkey. http://old.fedu.metu.edu.tr/ufbmek5/b_kitabi/PDF/Matematik/Bildiri/t219d.pdf

Dağlı, H., & Dağlıoğlı, H. E. (2017). Preschool teachers' pedagogical content knowledge about mathematics. In I. Koleva & G. Duman (Eds.), *Educational Research and Practice* (pp. 124-129). Sofia, Bulgaria: St. Kliment Ohridski University Press.

Dağlıoğlu, H. E., Dağlı, H., & Kılıç, N. M. (2014). Okul öncesi eğitimi öğretmen adaylarının matematik eğitimi dersine karşı tutumlarının çeşitli değişkenler açısından incelenmesi [Examination of pre-school teacher candidates' attitudes towards mathematics education course in terms of various variables]. In Proceedings of the *YILDIZ International Conference on Educational Research and Social Sciences Proceedings* (pp. 293-304). İstanbul, Turkey, Pegem Akademi.

Dağlıoğlu, H. E., Genç, H., & Dağlı, H. (2017). Gelişimsel açıdan okul öncesi dönemde matematik eğitimi. [Developmental aspects of mathematics education in preschool period] In İ. Ulutaş (Ed.), *Okul öncesi matematik eğitimi* [Preschool math education], (pp. 12-36). Ankara, Turkey: Hedef CS.

Erdoğan, S. (2006). *Altı yaş grubu çocuklarına drama yöntemi ile verilen matematik eğitiminin matematik yeteneğine etkisinin İncelenmesi* [A study on the effect of mathematıcs educatıon gıven wıth drama method to sıx-years-old chıldren on mathematıcs ability]. (Doctoral Dissertation). Ankara University, Ankara, Turkey. Retrieved from https://tez.yok.gov.tr

Even, R., & Tirosh, D. (1995). Subject-matter knowledge and knowledge about students as sourches of teacher presentations of the subject-matter. *Educational Studies in Mathematics, 29*(1), 1-20.

Fisher, P. H. (2004). *Early math interest and the development of math skills: an understudied relationship*. (Doctoral Dissertation). University of Massachusetts, USA.

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education, 2*(1), 79-96.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research, 79*(3), 1202-1242.

Gifford, S. (2005*). Teaching mathematics 3-5: developing learning in the foundation stage*. London, United Kingdom: Open University Press.

Gökkurt, B., & Soylu, Y. (2016). Ortaokul matematik öğretmenlerinin matematiksel alan bilgilerinin incelenmesi: Prizma örneği [Examination of mathematical knowledge of secondary school mathematics teachers: Prism example]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 16*(2), 451-481.

Güven, B., Karataş, İ., Öztürk, Y., Arslan, S., & Gürsoy, K. (2013). Okul öncesi öğretmenlerinin ve öğretmen adaylarının okul öncesi matematik eğitimine ilişkin inançların belirlenmesine yönelik bir ölçek geliştirme çalışması [A scale development study to determine the beliefs of preschool teachers and teacher candidates' about pre-school mathematics education]. *İlköğretim Online*, *12*(4), 969-980.

Güven, Y. (1998). Kız ve erkek çocuklarda matematik yeteneği ve matematik başarısı konusunda okulöncesi ve ilkokul (ilköğretim) öğretmenlerinin görüşlerinin değerlendirilmesi [Evaluation of preschool and primary school teachers' views on mathematics ability and mathematics achievement in girls and boys]. *Marmara Üniversitesi Atatürk Eğitim Fakültesi Eğitim Bilimleri Dergisi, 10*, 121-138.

Hacısalihoğlu Karadeniz, M. (2011). *Okul öncesi öğretmenlerinin sınıf içi matematik uygulamalarının okul öncesi eğitim programına uyumluluğu* [Accordance of preschool teachers' classroom mathematics practices with preschool education program]. (Doctoral Dissertation), Karadeniz Technical University, Trabzon, Turkey.

Henniger, M. L. (1987). Learning mathematics and science through play. *Childhood Education, 63*(3), 167-171.

Jang, Y. J. (2013) *Perspectives on mathematics education for young children* (Doctoral dissertation). University of Illinois, USA.

Karakuş, H., Akman, B., & Ergene, Ö. (2018). Matematiksel Gelişim İnanç Ölçeği'ni Türkçeye uyarlama çalışması [The Turkish Adaptation Study of the Mathematical Development Beliefs Scale]. *Eğitim ve Öğretim Dergisi, 8*(2), 211-228.

Karasar, N. (2012). *Bilimsel araştırma yöntemi* [Scientific research method]. Ankara, Turkey: Nobel.

Kilday, C. R. (2010). *Factors affecting children's math achievement scores in preschool* (Doctoral Dissertation). University of Virginia, USA.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Koç, F., Sak, R., & Kayri, M. (2015). Okul öncesi eğitim programındaki etkinliklere yönelik öz-yeterlik inanç ölçeğinin geçerlik ve güvenirlik analizi [Validity and reliability analysis of self-efficacy beliefs scale for activities in preschool education program], *İlköğretim Online*, *14*(4), 1416-1427. doi: http://dx.doi.org/10.17051/io.2015.50571

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: the role of model error. *Multivariate Behavioral Research, 36*(4), 611-637.

Masters, G. N. (1988). The analysis of partial credit scoring. *Applied Measurement in Education*, *1*(4), 279-297.

McCray, J. (2008). *Pedagogical content knowledge for preschool mathematics: relationships to teaching practices and child outcomes* (Doctoral Dissertation). Loyola University, Chicago, IL.

McCray J.S., & Chen,J.Q. (2012). Pedagogical content knowledge for preschool mathematics: construct validity of a new teacher ınterview, *Journal of Research in Childhood Education*, *26*(3), 291-307.

McGrath, C. (2010). *Supporting early mathematical development*. New York, NY: Routledge.

Metin, N. (1994). Okul öncesi dönemde matematik eğitimi ve etkinlik örnekleri [Math education and activity examples in preschool period]. Ş. Bilir (Ed.), *Okul öncesi eğitimciler için el kitabı* [Handbook for preschool educators]. Istanbul, Turkey: Ya-Pa.

Milli Eğitim Bakanlığı (Turkish Ministry of National Education), (2013). *Okul öncesi eğitim programı* [Preschool education program]. Ankara, Turkey: Milli Eğitim Bakanlığı.

Mokrova, I. L. (2012). *Motivation at preschool age and subsequent school success: role of supportive parenting and child temperament* (Doctoral Dissertation). University of North Carolina, USA.

Mononen, R., & Aunio, P. (2013). Early mathematical performance in Finnish kindergarten and grade one. *Lumat, 1*(3), 245-262.

Mononen, R., Aunio, P., & Koponen, T. (2014). Investigating right start mathematics kindergarten instruction in Finland. J*ournal of Early Childhood Education Research, 3*(1), 2-26.

National Association for the Education of Young Children & National Council of Teachers of Mathematics. (2010). *Position statement. Early childhood mathematics: Promoting good beginnings*. Retrieved from www.naeyc.org/resources/position_statements/psmath.htm

National Council of Teachers of Mathematics. (2000). *Principles and Standards for school mathematics.* Reston, VA: NCTM.

National Council of Teachers of Mathematics. (2009). *Where We Stand.* Reston, VA: NCTM. Retrieved from https://www.naeyc.org/files/naeyc/file/positions/ecmath.pdf

National Council of Teachers of Mathematics. (2013). *Mathematics in Early Childhood Learning—NCTM position.* Reston, VA: NCTM. Retrieved from http://www.nctm.org/uploadedFiles/Standards_and_Positions/Position_Statements/Early%20Childhood%20Mathematics%20(2013).pdf

Nicol, C., & Crespo, S. (2006). Learning to teach with mathematics textbooks: How pre-service teachers interpret and use curriculum materials. *Educational Studies in Mathematics, 62*(3), 331-355.

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw Hill.

Ontario Ministry of Education (OME). (2014). *Doing Mathematics with Your Child, Kindergarten to Grade 6: A Parent Guide*. Toronto, Ontario, Canada.

Parpucu, N., & Erdoğan, S. (2017). Okul öncesi öğretmenlerinin sınıf uygulamalarında matematik dilini kullanma sıklıkları ile pedagojik matematik içerik bilgileri arasındaki ilişki [The relationship between the frequency of mathematical language and pedagogical mathematic content knowledge of preschool teachers]. *Erken Çocukluk Çalışmaları Dergisi, 1*(1), 19-32. Doi: 10.24130/eccd.jecs.19672017118

Platas, L. M. (2008). *Measuring teachers' knowledge of early mathematical development and their beliefs about mathematics teaching and learning in the preschool classroom* (Doctoral Dissertation). University of California, USA.

Sammons, P., Elliot, K., Sylva, K., Melhuish, E., Siraj-Blatchford, I., & Taggart, B. (2004). The impact of pre-school on young children's cognitive attainment at entry to reception. *British Educational Research Journal*, *30*(5), 691-712.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Research*, *15*(2), 4-14.

Şeker, P. T., & Alisinanoğlu, F. (2015). A survey study of the effects of preschool teachers' beliefs and self-efficacy towards mathematics education and their demographic features on 48-60 month-old preschool children's mathematic skills. *Creative Education*, *6*(3), 405-414. doi: 10.4236/ce.2015.63040

Tanışlı, D. (2013). İlköğretim matematik öğretmeni adaylarının pedagojik alan bilgisi bağlamında sorgulama becerileri ve öğrenci bilgileri [Preservice primary school mathematics teachers' questioning skills and knowledge of students in terms of pedagogical content knowledge]. *Eğitim ve Bilim, 38*(169), 80-95.

Tella, A. (2007). The impact of motivation on student's academic achievement and learning outcomes in mathematics among secondary school students in Nigeria. *Eurasia Journal of Mathematics, Science & Technology Education, 3*(2), 149-156.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson.

Türkiye İstatistik Kurumu Başkanlığı [Turkish Statistical Institute] (TÜİK). (2012). *Gelir ve yaşam koşulları araştırması mikro veri seti* (Kesit) [Income and living conditions research micro data set (Section)]. Ankara, Turkey: TÜİK. Retrieved from http://www.tuik.gov.tr/MicroVeri/GYKA_2012/turkce/metaveri/siiniiflamalar/index.html

Umay, A. (2003). Matematiksel muhakeme yeteneği [Mathematıcal Reasonıng Ability]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 24*, 234-243.

Watts, T. W., Duncan, G. J., Clements, D. H., & Sarama, J. (2017). What is the long-run impact of learning mathematics during preschool? *Child Development*, *89*(2), 539-555. DOI: 10.1111/cdev.12713

Yurdugül, H. (2005). *Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması* [Using scope validity indices for scope validity in scale development studies]. In Proceedings of the Pamukkale University, XIV. National Educational Sciences Congress, Pamukkale University, Denizli, Turkey.

Zhang, Y. (2015) *Pedagogical content knowledge in early mathematics: what teachers know and how it associates with teaching and learning* (Doctoral dissertation). Loyola University, Chicago, IL.

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 636–655
https://dx.doi.org/10.21449/ijate.623080

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

Using the Malmquist Index in Evaluation Process to Enhance Mathematical Literacy in High School Students

Niusha Mostoli [1], **Mohsen Rostamy** [1,*], **Ahmad Shahverani** [1], **Mohammad Hasan Behzadi** [1]

[1]Department of Mathematics, Science and Research Branch, Islamic Azad University, Tehran, Iran

ARTICLE HISTORY

Received: 21 September 2019
Revised: 22 November 2019
Accepted: 17 December 2019

KEYWORDS

Productivity Evaluation,
The Malmquist Index,
Mathematical Literacy,
Evaluation instrument
GAMS,

Abstract: The study aimed to calculate the evaluation of 9th grade female students and compare the development of the educational process in increasing mathematical literacy using the Malmquist Index at different time intervals. This educational process was accomplished by analysing and integrating realistic mathematical education and mathematical problem-solving. The populations of the study were 120 ninth grade female students. Each student was as a DMU whose inputs were the math test score and the outputs were math test score of December and June. The data analysis method was based on (DEA) technique to calculate efficiency. The output-driven (CCR) model was used to determine students' performance coefficient. Then, the Malmquist Index was used to compare productivity evaluation after the end of the training course in December and the end of the year in June. In general, the results from changes in productivity evaluation of students using the Malmquist Index showed that the students in the experimental group who learned problem-solving and realistic mathematics had an increase in the overall productivity evaluation factor after completing the training compared to the others.

1. INTRODUCTION

We live in a developing world where if we do not nurture students with an efficient and dynamic education system, we will be left out of world educational standards. That’s why education has long been of particular interest to mankind. Today, in modern societies, statesmen believe that an integrated curriculum must be used to advance and achieve the prescribed goals. One of the biggest challenges in writing an integrated curriculum at any level is to make the complex concepts understandable at the same time with preserving their integrity for students (Fendel, 2012). Learning math is different from learning other subjects for different reasons. Mathematics is the language of nature's explanation and is based on reason and creativity. Besides examining the targeted learning content, the use of modern methods in evaluating the educational methods is another concern of the educational system of any country. Education system plays an essential role in economic and social development of any country because of its mission in raising the required expert manpower. Therefore, assessing the performance of its different domains is of great importance (Stacey et al., 2015). In addition to satisfying

CONTACT: Mohsen Rostamy ✉ Mohsen_rostamy@yahoo.com 🖃 Department of Mathematics, Science and Research Branch, Islamic Azad University, Tehran, Iran

ISSN-e: 2148-7456 /© IJATE 2019

personal interests, mathematics has also been studied for practical objectives in other fields of study (Gatabi, Stacey, & Gooya, 2012). In confirming the effective communication of mathematics with other scientific fields, Gauss considered the math as the queen of sciences (Gatabi et al., 2012).

A closer look can show the effect of math skills on different levels of life. Hence, it can be said that student math success is always dependent on the future of a country, so the desire to understand and identify the factors leading to student math success, it has always been crucial to national leaders, policymakers, and educators (Burnett, 2005). According to Freudenthal, math must be taught for its usefulness, and of course this is not achieved through useful mathematical teaching, since any subject of mathematics can, however, be useful in limited fields (Gravemeijer & Terwel, 2000). One of these useful methods is problem-solving and combining it with the real world. Problem-solving is a vital skill for living in the present age. Nowadays, authorities are called for high-level thinking skills and problem-solving, both in the public and in the field of technologies in all activities (Stacey et al., 2015).

Given the importance of mathematics training and the stated goals, it is clear that the real-world problem-solving ability is one of the issues emphasized and endorsed by policy makers in the education system. One way to determine the realization of these goals are identification and participation in international tests. Programme for International Student Assessment (PISA) is an international study that emphasizes the application of mathematics to everyday life. The main question of the PISA study of Organization for Economic Cooperation and Development on mathematics is if students are mathematically prepared for future challenges (Adams & Wu, 2000). The PISA study has emerged to answer the assessment of 15-year-old students' readiness to address future challenges in after-school life, not just school life. This study has been conducted every three years since 2000 (OECD, 2002). The tests used in PISA studies include issues that measure students' ability to cope with real-world challenges, taking into account different criteria (Development, n.d.)

1.1. In framework of the PISA study, mathematical literacy is defined as follows:

Mathematical literacy is an individual's talent to formulate, apply, and interpret mathematics in a variety of fields, including mathematical reasoning and the use of mathematical concepts, methods, facts, and tools to describe, express, and predict phenomena. Mathematics literacy helps people to understand the role of mathematics in the world and to make the reasonable judgments and decisions needed for a productive, committed and thoughtful citizen (Alvarez, 2018).

Real-world applications of the curriculum are serious challenges in targeting the school curriculum. Given recent targeting in education organization, it is important to address students' ability to apply mathematics in everyday life. A topic of interest today is the gap between the mathematical world and the real world, resulting in students' inability to use mathematics in the real world (Alimohammadlou & Mohammadi, 2016). According to the latest National Curriculum Document, students' empowerment in applying mathematics to solve everyday problems and abstractions are one of the main goals of mathematical education in the education system (Gravemeijer, 1994). In addition, the Higher Education Council in the first high school goals approval emphasizes that first-grade high school students must be proficient in the use of mathematics to solve problems for themselves and society at the end of the course. Mathematics education should therefore provide an opportunity for students to experience the relationship between the real world and the mathematical world, thereby solving their everyday problems (Breen, Cleary, & Shea, n.d.).

It can now be argued that the education system undoubtedly needs performance evaluation and assessment in order to make the best use of its limited resources and better effectiveness. One of the most effective tools in this field is data envelopment analysis which is used as a non-parametric method to calculate the efficiency of decision making units (Charnes, Cooper, & Rhodes, 1978). Today, DEA technique is rapidly expanding and is being used in the evaluation of various organizations and industries such as the banking industry, post offices, and hospitals, training canters, power plants, refineries and more (Wang & Lan, 2011). Various papers and studies have been presented to date based on this technique, which are mainly based on conventional DEA models, such as CCR, BCC, etc. These models are not capable of evaluating the performance of multiple-component decision making units (F äre, Grosskopf, & Roos, 1995). Management needs methods to do so because of the need to know the performance of system components. Therefore, in 1989, scientists such as Fare, Grosskopf, Lindgren and Roos used the data envelopment analysis technique to calculate the Malmquist Index. In 1992, the Malmquist Index was divided into two factors, efficiency changes, technology changes, and changes in technology by the scientists. Every effort to increase efficiency and productivity, which involves measuring, analysing, planning and improving productivity, falls into the productivity cycle, and measuring productivity evaluation is the first and foremost task in this cycle (Fare, Grosskopf, Lindgren & Roos, 1992).

1.2. In this regard, the present study seeks to answer the following questions:

- Does a teaching problem-solving technique help students to change, apply and interpret the relationship of content areas to problems and find solutions?
- Does a teaching realistic math technique help students apply math textbook problems and model simple real-life situations?
- Are the Malmquist and the GAMS effective tools in evaluating the performance of 9th grade students in mathematics literacy test?
- Is it possible to evaluate and compare students' performance using the Malmquist Index and Data Envelopment Analysis?

1.3. Conceptual definitions

Data Envelopment Analysis: It is a mathematical programming-based method that enables the calculation of technical performance and Evaluate the data for desired programs with some inputs and outputs of decision making units (DMU) without assigning weights to inputs and outputs and matching them (Sağlam, 2017).

The output-driven CCR model: The name of this model (CCR) is derived from the first letters of the three scholars, namely, Charles, Cooper, and Rhodes. The model has constant-scale output. The output models seek to maximize outputs without any increase (change or decrease) in the amount of inputs(Sinuany-Stern, Mehrez, & Barboy, 1994). The purpose of this mode is to maximize the output without increasing inputs or resources. The model is shown in equation 1:

$$
\begin{aligned}
&\text{Max } \varphi \qquad (1)\\
&\text{s.t}\\
&\sum_{j=1}^{n} \gamma_j x_{ij} \leq x_{io}, \qquad i = 1, \ldots,\\
&\sum_{j=1}^{n} \gamma_j\, y_{rj} \geq \varphi y_{ro}, \qquad r = 1, \ldots, s\\
&\gamma_j \geq 0\,, \qquad j = 1, \ldots, n
\end{aligned}
$$

The model is always feasible and the optimal solution applies to the condition φ^*≥1. If φ^*=1 then the DMU is technically efficient in the nature of the output. If φ^*>1 then the DMU is inefficient in the nature of the output (Basic & Model, 2005). Using this data envelopment analysis model to calculate the efficiency and rankings of units, more than one unit may achieve the highest efficiency coefficient, i.e. 1. In this case, it is not possible to compare and rank these efficient units. In this case, the Anderson-Peterson (AP) method can be used to rank efficient units (Charnes et al., 1978).

Anderson and Petersen Ranking (AP): Anderson and Petersen (1993) have developed a method that is suitable for ranking efficient units and can help to compare and contrast units that have performance 1. The method in linear programming model for DMU with efficiency 1, smaller constraint and equal to zero; corresponding to DMU is eliminated so that DMU does not face resource constraint (Hosseinzadeh Lotfi et al., 2013). Then the model is resolved after the changes are made. In this case the efficiency coefficient of the efficient units may be larger, the unit is more efficient (Andersen & Petersen, 1993).

Productivity: The performance of a company in converting the input to the output can be expressed in a variety of ways. One of the ways to measure performance is productivity ratio. By defining a firm's productivity as the output-to-input ratios, values larger than this ratio reflect the firm's better performance. The profitability is a relative relationship. Productivity is the efficiency in using measured resources as the output relative to input. To calculate productivity, it is necessary to calculate the input and output values. Productivity is technically broken down into two factors: efficiency and effectiveness (Tohidi & Razavyan, 2013).

Efficiency: It is the ratio of actual yields to standardized and expected yields of efficiency, or in fact the ratio of the amount of work done to the amount of work to be done (Emrouznejad & Yang, 2018).

Malmquist: The Malmquist Productivity Evaluation Index is a two-way Index that calculates productivity growth between two units (firm) in one period, or one firm in two different periods. Stan Malmquist, Swedish economist, (1953) introduced the Malmquist Index as the standard of living, and in 1982, it was first applied to production theory by Christensen and Diewert (Caves, Christensen & Diewert, 2012). In 1989, Farr et al., used data envelopment analysis to calculate the Malmquist Index, and in 1994, the Index was broken down into two factors: efficiency and technology (Balf, Lotfi, & Alizadeh, 2010). The data computed for the distance functions are the technical efficiency obtained from data envelopment analysis equations. Thus, the Malmquist Productivity Index is defined by the maximization between the two times t and t + 1, with respect to the common efficiency boundary at time t as equation 2.

$$\mathbf{M_0^t}\,(\mathbf{Y_t},\mathbf{X_t}\,,\mathbf{Y_{t+1}}\,,\mathbf{X_{t+1}}) = \frac{\mathbf{d_0^t}\,(\mathbf{Y_{t+1}},\mathbf{X_{t+1}})}{\mathbf{d_0^t}(\mathbf{Y_t}\,,\mathbf{X_t})} \tag{2}$$

Similarly, the Malmquist Productivity Index is defined by the maximization between the two times t and t + 1 with respect to the current efficiency boundary at time t + 1 as the equation 3.

$$\mathbf{M_0^{t+1}}(\mathbf{Y_t},\mathbf{X_t}\,,\mathbf{Y_{t+1}}\,,\mathbf{X_{t+1}}) = \frac{\mathbf{d_0^{t+1}}(\mathbf{Y_{t+1}},\mathbf{X_{t+1}})}{\boldsymbol{d_0^{t+1}(Y_t\,,X_t)}} \tag{3}$$

The two Malmquist indices are equivalents, and the Malmquist productivity change Index is expressed as the geometric mean of the two productivity indices and can be represented by the equation 4.

$$\mathbf{M_0}(\mathbf{Y_t},\mathbf{X_t}\,,\mathbf{Y_{t+1}}\,,\mathbf{X_{t+1}}) = \left[\frac{\mathbf{d_0^{t+1}}(\mathbf{Y_{t+1}},\mathbf{X_{t+1}})}{\mathbf{d_0^{t+1}}(\mathbf{Y_t}\,,\mathbf{X_t})} \times \frac{\mathbf{d_0^t}\,(\mathbf{Y_{t+1}},\mathbf{X_{t+1}})}{\mathbf{d_0^t}(\mathbf{Y_t}\,,\mathbf{X_t})}\right]^{\frac{1}{2}} \tag{4}$$

This productivity equation expresses the point (Y $_{(t+1)}$, X $_{(t+1)}$) versus point (Y $_t$, X $_t$). Values greater than one represent productivity growth. If performance declines during the trend, the Malmquist Index will be less than 1. Equation 4 can be broken down into equation 5 to allow the Malmquist Index to show technological change, production scale and technical efficiency (Wang & Lan, 2011).

$$\mathbf{M_0}(\mathbf{Y_t}, \mathbf{X_t}, \mathbf{Y_{t+1}}, \mathbf{X_{t+1}}) = \frac{\mathbf{d_0^{t+1}}(\mathbf{Y_{t+1}}, \mathbf{X_{t+1}})}{\mathbf{d_0^{t}}(\mathbf{Y_t}, \mathbf{X_t})} \left[\frac{\mathbf{d_0^{t}}(\mathbf{Y_{t+1}}, \mathbf{X_{t+1}})}{\mathbf{d_0^{t+1}}(\mathbf{Y_{t+1}}, \mathbf{X_{t+1}})} \times \frac{\mathbf{d_0^{t}}(\mathbf{Y_t}, \mathbf{X_t})}{\mathbf{d_0^{t+1}}(\mathbf{Y_t}, \mathbf{X_t})}\right]^{\frac{1}{2}} \qquad (5)$$

In equation 5, the term outside the bracket represents the change in the technical efficiency at t and t + 1 and equals to the ratio of the technical efficiency at time t + 1 to the technical efficiency at time t. The term inside the bracket represents the technological shift between the two above times. M_0 greater than 1 indicates that productivity evaluation has increased between the two periods. This increase can be based on technical efficiency or technology advancement (change of efficient border) (Diewert & Fox, 2010).

Realistic Mathematical Education: In the late 70s, Freudenthal et al., objected to the American movement and the mechanical mathematics education approach in the Netherlands, and explained the theory of realistic mathematics education (the new mathematics) to reform the process of teaching and learning mathematics (Lange, 1987). The underlying philosophy of realistic mathematic education was that the learner needed to gain mathematical understanding by working on the fields which was meaningful for him (Freudenthal, 1973). In his view, the real world is the source or starting pint of mathematic concept development (Koyuncu, Guzeller, & Akyuz, 2016). Mathematics education in this way is guided by reinvention so that the student can experience it himself (Esther, P érez, Duque, & Garc ía, 2018). The three key principles of this approach are Guided reinvention, didactical phenomenology, and self-developed model (Sumirattana, Makanong, & Thipkong, 2017).

Problem-Solving: From George Polya’s perspective, problem-solving strategy is done in the following four steps (Pólya, 1962):

1. **(Understanding the problem**) what is required in the problem?
2. **(Deeper recognition of problem and map design**) how the different components of the problem are interconnected and what is the missing link to the problem data?
3. **(Implementation of map for problem-solving**) this step depends on the correct implementation of steps 1 and 2. In fact, the major task to solve the problem is to get an idea of what the map is and how it works.
4. **(Review and getting back**) controlling the correct execution of the map (Esther et al., 2018).

The problem-solving process in this research is based on the Polya's mathematical model and the Shewhart cycle consisting of five parts that including Definition D, Assessment A, plan P, Implementation I, communication C, (Sumirattana et al., 2017).

The PISA Study: The first PISA study was conducted in 2000. The study was held every three years and the results of each course were published by the Organization for Economic Cooperation and Development. The main question of the PISA about mathematics is whether 15-year-old students are mathematically prepared for future challenges in after-school life. Testing this study focuses on real-world mathematics and operates beyond the school matters(Programme for International Student Assessment & OECD 2009).

Evaluation Instrument GAMS: software is a powerful and comprehensive tool for solving mathematical models even in large dimensions and disciplines of science, that is, wherever it is necessary to make optimal decisions with time, cost, and resources, the mathematical modelling should be used and GAMS is a highly efficient tool for solving these types of models. The most important application of GAMS is the optimization of research models in operations and data envelopment analysis and data evaluate (Rosenthal, 2007).

1.4. Research history

One of the researches related to mathematical literacy assessment in the world is a research in which a test is designed to fit the mathematical knowledge level of 15 years old students. Based on the results of this study, the mathematical literacy of these students is reported in the field of understanding undesirable problem and in the field of interpreting, using processes, modelling and very undesirable describing (Sari & Valentino, 2016).

In another study in Turkey, students' success rates in answering a variety of the PISA test questions in 2003 and 2012 were compared. According to this study, students have been more successful in the both years in answering multiple-choice questions (Thomson, Hillman, & Bortoli, 2013).

In another study, students' performance in dealing with new problems but commensurate with their mathematical knowledge was measured by a PISA test. The results of this study suggest that math literacy among Irish 18- and 19-year-old students has been at a desirable level due to their education in engineering fields, which is expected because of mathematical training (Breen, Cleary, & Shea, n.d.)

Koçak, Türe and Atan (2019) in another study did researches Efficiency Measurement with Network DEA. They did compute the educational economy efficiency of the Organisation for Economic Co-operation and Development. They study is the use of a novel approach to computing the educational economic efficiency using relational network DEAL with GAMS (Koçak et al, 2019).

Chen and Yao (2010) calculated the Malmquist Index for measuring total productivity changes and its components in three important industries of China, including textile, chemical and metal industries during four development plans of China by using Data Envelopment Analysis Method. The results showed that productivity changes in 1966 in the fourth and fifth programs compared to the previous programs, but in the sixth plan slightly increased (about 3 to 5%). Total productivity components didn't change significantly during programs (Chen & Lin, n.d.).

Wei and Hao (2011) studied the role of human capital on the total productivity growth of factors in 30 Chinese provinces during 1985-2004 and used the Malmquist Index to measure productivity growth. Results indicated that human capital had a positive significant effect on the total productivity growth (Wei & Hao, n.d.).

Maodus et al., (1998) examined the effect of human capital on productivity in OECD countries during 1965-1990. They used data envelopment analysis to perform this study. Results showed that higher level of human capital increases productivity (Maudos & Pastor, 1998).

Chen and Lin (2006) conducted a case study on the R&D performance of 52 integrated semiconductor companies located at the Sino Chou Science and Technology Park in Taiwan using the DEA approach. Using the BCC model, they calculated the rates of technical and scale efficiency. Results showed that R&D performance is very different among the firms evaluated, and many inefficient firms have to increase their economic scales (Chen & Lin, n.d.).

2. METHOD

2.1. Variables

Accurate and appropriate selection of inputs and outputs is one of the determinants of achieving reliable and proportional outcomes for educational purposes in order to calculate correct values of productivity in different periods. In this study, each student is considered as a DMU with two approaches to realistic mathematics education and the problem-solving process. Input of each DMU (student) includes two realistic math instruction and math problem-solving methods. The outputs in each DMU includes performance of each student using a mathematical literacy performance assessment taking into account the score obtained in pre-test in the beginning of the academic year M and the mathematical post-test will be at the end of the first three months A and the final exam at the end of academic year P. Finally, the productivity was calculated using the Malmquist Index in the A-M and A-P intervals. GAMS software was used to calculate the efficiency and productivity evaluation and finally the efficient units were sorted by the Anderson and Peterson AP ranking method.

2.2. Methods

The statistical population consisted of the 9th grade female students of District 3 of Tehran. The sample was selected by simple random sampling due to the high size of the statistical population. 120 students were selected as the experimental group, dividing into 4 classes based on Table 1.

Table 1. Classes Model

Trained education methods	Number	Class
Just problem-solving (PS)	30	A
Just realistic mathematical method (RME)	30	B
Both educational methods (PS & RME)	30	C
Without educational method	30	D

Before performing the research, students were given a mathematical pre-test that questions were calculated similar to post-test with validity and reliability (r = 0.209 - 0.743, p = 0.243 - 0.569) and Cronbach's alpha 0.754 The math pre-test was conducted at the beginning of the M academic year. After doing the research to evaluate the students' mathematical literacy and calculating the final efficiency of questions, math post-test was conducted based on the PISA (2015) at the end of our first three academic years A. The questions in this test were consistent with realistic mathematics content and in accordance with the objectives of mathematical problem-solving. First, students were given mathematical literacy pre-test, and then students of experimental group were taught problem-solving and realistic mathematics in 12 weeks. Finally, the mathematics test at the end of academic year P was conducted in June. After localization, the final test was given to the experts to criticize the publishable questions in the PISA Studies 2012. After making corrections and approving by the experts, taking into account the mathematical knowledge of 15-year-old students, a number of questions that did not fit with the topics of mathematics books being taught in Iran were eliminated. Finally, a test consisting of 10 questions with maximum similarity to the mathematical test of PISA was prepared. The questions were given to several mathematics professors and mathematics educators and some experienced high school math teachers and approved after review. Based on primary study, Cronbach's alpha coefficient of this test was 0.73. Given that this value was greater than 0.7, it showed good reliability. As mentioned earlier, students were taught problem-solving and realistic mathematics methods in 12 weeks:

First three weeks of the course: Students were taught that the following should be clear to solve a problem: What to find? What is the unknown? What is the assumption? What is the known? What is the relationship between known and unknown? Students review the content of the problem; the teacher should guide them based on their request and teach them how to use and apply mathematical definitions. Students need to evaluate problem situations to provide meaningful, simple models for problem-solving.

Second three weeks of the course: Students were taught to draw a problem-solving map using tables, figures, diagrams, and data, so they could figure out a correct way to solve it.

Third three weeks of the course: The students were taught to apply all their guesses and plans until they felt that they might not solve the problem, so they could prepare and execute a new project.

Final three weeks of the course: The student was finally taught to review all the steps taken, re-examine their arguments and answers, and analyse the correctness of their solution.

Notice that throughout the courses, inspired by realistic mathematics, students were given the opportunity to create their own mathematical knowledge, invent new mathematics, and relate abstract mathematics to the real world. Given that each student is considered as a DMU, the performance of each student in the first stage was calculated with the first test score in October and the second test score in December EMA and in the second stage with the second test score in December and score the final test in June EAP. The total score of each test is 20, i.e. 10 questions with 2 points each. The performance of the data was calculated using the output-driven CCR model. At the end of the year after the calculation of final performance, the Malmquist Index was used to calculate data productivity evaluation. The research model is presented in Figure 1.

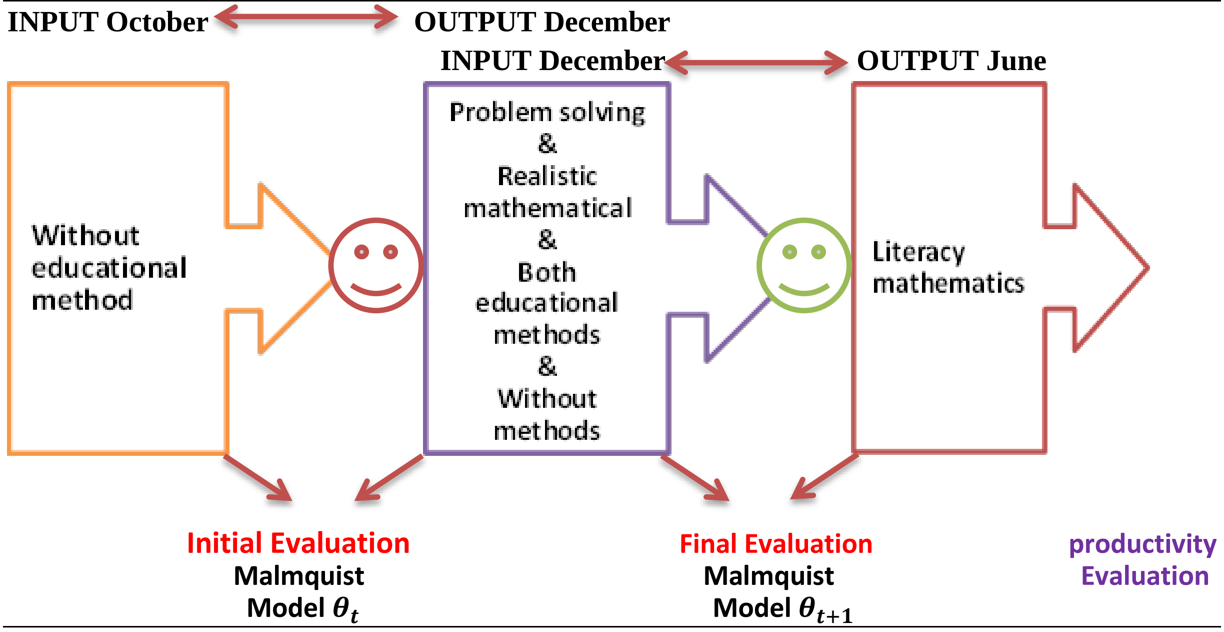


Figure 1. Input & Output Model of the Malmquist Index

3. RESULT

In this section, considering the results of tests taken by the students, data will be described and the questions answered. Table 2 briefly refers to the used terms.

Table 2. Abbreviation Table

Abbreviation	Explanation	Abbreviation	Explanation
DMU	Decision Making Unit	TCI	Technical Change Index
RME	Realistic Mathematics Education	SECI	Scale Efficiency Change Index
PS	Problem Solving	FGLR	Fare, Grosskof, Lindgren and Roos
EMA	Efficiency October-December	FGLR	Malmquist Model
EAP	Efficiency December- June	FPCI	Factory Productivity Change Index

Question 1: Does teaching problem-solving techniques and applying and interpreting change the content contexts of problems and finding solutions, help students?

Inputs are Class A who received math problem-solving training and the output is enhanced math literacy according to Table 3. The mean percent of performance in the initial test before problem-solving method was 23.3% and at the end of training 36.6% and the mean productivity with the Malmquist Index was 63.3%. At a glance, according to Table 3, it can be said that students' mathematical performance after receiving problem-solving, had better results for Class A. Although the number of students who became effective after receiving the problem-solving instruction was not high, the math scores with the Malmquist mean score showed improvement in the final math test.

Table 3. Problem-Solving Efficiency

CLASS A							
ROW	**DMU**	**E(MA)**	**E(AP)**	**TCI**	**SECI**	**FGLR**	**FPCI**
1	DMU1	0.807	1.002	0.998	1.232	1.229	Increase
2	DMU2	1	0.875	0.947	0.875	0.829	Decrease
3	DMU3	0.969	0.797	1.193	0.823	0.982	Decrease
4	DMU4	0.81	0.719	0.984	0.888	0.874	Decrease
5	DMU5	1	1	0.074	1	0.974	Decrease
6	DMU6	0.722	0.733	0.975	1.044	1.018	Increase
7	DMU7	0.954	0.906	0.907	0.905	0.922	Decrease
8	DMU8	1	1	0.939	1	0.939	Decrease
9	DMU9	0.715	0.784	0.98	1.098	1.076	Increase
10	DMU10	0.392	1	0.971	1.038	1.007	Increase
11	DMU11	1	1	1.640	1	1.640	Increase
12	DMU12	0.665	1.01	1.201	1.177	1.413	Increase
13	DMU13	0.734	0.704	1.520	1	1.459	Increase
14	DMU14	0	0	0	0	0	0
15	DMU15	0.809	0.993	0.996	1.234	1.229	Increase
16	DMU16	0.39	1	0.971	1.038	1.007	Increase
17	DMU17	0.715	0.784	0.98	1.098	1.077	Increase
18	DMU18	0.705	0.794	0.96	1.097	1.076	Increase
19	DMU19	0.382	0.417	0.97	1.039	1.007	Increase
20	DMU20	1	1	0.074	1	0.974	Decrease
21	DMU21	0.709	0.79	0.98	1.098	1.076	Increase
22	DMU22	1	1	0.939	1	0.939	Decrease
23	DMU23	0.725	0.73	0.975	1.044	1.018	Increase
24	DMU24	0.704	0.795	0.96	1.097	1.076	Increase
25	DMU25	0.725	1	0.975	1.044	1.018	Increase
26	DMU26	0.959	0.901	0.907	0.905	0.922	Decrease
27	DMU27	0.49	1.03	0.971	1.038	1.007	Increase
28	DMU28	1	0.875	0.947	0.875	0.829	Decrease
29	DMU29	0.715	0.83	0.97	1.049	1.018	Increase
30	DMU30	0.665	0.782	1.201	1.177	1.413	Increase

Question 2: Does teaching realistic mathematics help students use mathematical textbook problems and model simple life situations?

Inputs are Class B who receives realistic mathematics instruction and the output is enhanced mathematical literacy according to Table 4. The mean percent of performance in the initial test before realistic mathematics test was 20% and at the end of the training period was 33.3% and the mean productivity with the Malmquist Index was 53.3% at a glance, according to Table 4, it can be said that students' mathematical performance after performing realistic mathematics had better results for class B.

Table 4. RME Efficiency

CLASS B							
ROW	**DMU**	**E(MA)**	**E(AP)**	**TCI**	**SECI**	**FGLR**	**FPCI**
1	DMU1	0.875	1	0.947	0.875	0.828	Decrease
2	DMU2	0.875	1	0.947	0.875	0.829	Decrease
3	DMU3	0.797	0.969	1.193	0.823	0.983	Decrease
4	DMU4	0.719	0.81	0.984	0.888	0.874	Decrease
5	DMU5	1	1	0.074	1	0.974	Decrease
6	DMU6	0.733	0.722	0.975	1.044	1.018	Increase
7	DMU7	0.906	0.954	0.907	0.905	0.922	Decrease
8	DMU8	1	1	0.939	1	0.939	Decrease
9	DMU9	0.784	0.715	0.98	1.098	1.076	Increase
10	DMU10	0.407	0.392	0.971	1.038	1.007	Increase
11	DMU11	1	1	1.640	1	1.640	Increase
12	DMU12	0.87	1	0.952	0.875	0.829	Decrease
13	DMU13	0.704	0.734	1.520	1	1.459	Increase
14	DMU14	0.779	0.72	0.98	1.098	1.067	Increase
15	DMU15	0.993	0.809	0.996	1.234	1.227	Increase
16	DMU16	0.409	0.39	0.971	1.038	1.008	Increase
17	DMU17	0.784	0.715	0.98	1.098	1.077	Increase
18	DMU18	1	1	0.939	1	0.939	Decrease
19	DMU19	0.417	0.382	0.97	1.039	1.009	Increase
20	DMU20	1	1	0.074	1	0.974	Decrease
21	DMU21	0.79	0.709	0.98	1.098	1.076	Increase
22	DMU22	1	1	0.939	1	0.939	Decrease
23	DMU23	0.721	0.79	0.984	0.888	0.874	Decrease
24	DMU24	0.795	0.704	0.96	1.097	1.076	Increase
25	DMU25	0.73	0.725	0.975	1.044	1.018	Increase
26	DMU26	0.901	0.959	0.907	0.905	0.923	Decrease
27	DMU27	0.399	0.49	0.971	1.038	1.007	Increase
28	DMU28	0.875	1	0.947	0.875	0.829	Decrease
29	DMU29	0.83	0.715	0.97	1.049	1.018	Increase
30	DMU30	0.812	0.635	1.201	1.177	1.412	Increase

Question 3: Are Malmquist and GAMS effective tools in evaluating the performance of 9th grade students in mathematics literacy test?

Inputs are Class C who receives realistic mathematics instruction and problem-solving method and the output is enhanced mathematical literacy according to Table 5. The mean percent of performance in the initial test before realistic mathematics test was 13.3% and at the end of the training period was 46.6% and the mean productivity with the Malmquist Index was 80%. At a glance, according to Table 5, it can be said that students' mathematical performance after performing realistic mathematics and problem-solving had better results for Class C.

Table 5. Both Educational Methods Efficiency

CLASS C							
ROW	**DMU**	**E(MA)**	**E(AP)**	**TCI**	**SECI**	**FGLR**	**FPCI**
1	DMU1	0.722	0.733	0.975	1.044	1.018	Increase
2	DMU2	1	0.875	0.947	0.875	0.829	Decrease
3	DMU3	0.715	1.002	0.97	1.049	1.019	Increase
4	DMU4	0.633	1	1.201	1.177	1.413	Increase
5	DMU5	0.715	1.01	0.98	1.098	1.075	Increase
6	DMU6	0.722	1.05	0.975	1.044	1.018	Increase
7	DMU7	0.954	0.906	0.907	0.905	0.922	Decrease
8	DMU8	1	1.004	0.947	0.875	0.829	Decrease
9	DMU9	0.715	0.784	0.98	1.098	1.076	Increase
10	DMU10	0.392	0.407	0.97	1.035	1.002	Increase
11	DMU11	0.49	0.399	0.971	1.038	1.007	Increase
12	DMU12	0.725	1.002	0.98	1.098	1.077	Increase
13	DMU13	0.665	0.782	1.201	1.177	1.413	Increase
14	DMU14	0.72	0.779	0.98	1.098	1.067	Increase
15	DMU15	0.799	1.2	0.996	1.234	1.227	Increase
16	DMU16	0.39	1.03	0.971	1.038	1.018	Increase
17	DMU17	0.715	0.784	0.98	1.098	1.075	Increase
18	DMU18	0.704	1.01	0.96	1.097	1.076	Increase
19	DMU19	1	1	1.640	1	1.640	Increase
20	DMU20	0.725	0.73	0.975	1.044	1.018	Increase
21	DMU21	0.709	1.11	0.98	1.098	1.076	Increase
22	DMU22	1	0.999	0.939	1	0.969	Decrease
23	DMU23	0.79	0.721	0.984	0.888	0.874	Decrease
24	DMU24	0.704	0.795	0.96	1.097	1.076	Increase
25	DMU25	0.725	0.73	0.975	1.044	1.018	Increase
26	DMU26	0.959	0.901	0.907	0.905	0.913	Decrease
27	DMU27	0.49	0.399	0.971	1.038	1.007	Increase
28	DMU28	0.807	1.22	0.998	1.232	1.229	Increase
29	DMU29	0.715	1.1	0.97	1.049	1.018	Increase
30	DMU30	0.625	1.03	1.201	1.177	1.410	Increase

Class D participated in the tests without any training. According to Table 6, the mean percent of performance in the initial test was 33.3% and at the end of the first three month of experimental group was 20% and the total mean productivity with the Malmquist Index was 20%. At a glance, according to Table 6, it can be said that students' mathematical performance without receiving realistic mathematics methods and problem-solving, had unfavourable results compared to the beginning of the academic year for Class D.

According to the Tables, the use of the Malmquist Index facilitates the achievement of students in various time periods in the academic year. The Malmquist and the GAMS are used to assess the quality of learning performances of 9th grade.

Table 6. Without Educational Method Efficiency

CLASSD							
ROW	**DMU**	**E(MA)**	**E(AP)**	**TCI**	**SECI**	**FGLR**	**FPCI**
1	DMU1	1	0.875	0.947	0.875	0.829	Decrease
2	DMU2	0.715	0.784	0.98	1.098	1.077	Increase
3	DMU3	1	1	0.939	1	0.939	Decrease
4	DMU4	0.382	0.417	0.97	1.039	1.009	Increase
5	DMU5	0.969	0.797	1.193	0.823	0.982	Decrease
6	DMU6	0.81	0.719	0.984	0.888	0.874	Decrease
7	DMU7	1	1	0.939	1	0.939	Decrease
8	DMU8	0.79	0.721	0.984	0.888	0.874	Decrease
9	DMU9	0.969	0.797	1.193	0.823	0.983	Decrease
10	DMU10	0.725	0.73	0.975	1.044	1.018	Increase
11	DMU11	0.959	0.901	0.907	0.905	0.923	Decrease
12	DMU12	0.49	0.399	0.971	1.038	1.007	Increase
13	DMU13	1	0.875	0.947	0.875	0.829	Decrease
14	DMU14	0.969	0.797	1.193	0.823	0.982	Decrease
15	DMU15	0.81	0.719	0.984	0.888	0.874	Decrease
16	DMU16	0.797	0.969	1.193	0.823	0.983	Decrease
17	DMU17	0.719	0.81	0.984	0.888	0.874	Decrease
18	DMU18	0.969	0.797	1.193	0.823	0.983	Decrease
19	DMU19	0.81	0.719	0.984	0.888	0.874	Decrease
20	DMU20	1	1	0.074	1	0.974	Decrease
21	DMU21	0.959	0.901	0.907	0.905	0.922	Decrease
22	DMU22	0.954	0.906	0.907	0.905	0.922	Decrease
23	DMU23	1	1	0.939	1	0.939	Decrease
24	DMU24	0.715	0.784	0.98	1.098	1.076	Increase
25	DMU25	1	0.999	0.939	1	0.969	Decrease
26	DMU26	1	1	1.640	1	1.640	Increase
27	DMU27	1	0.87	0.952	0.875	0.829	Decrease
28	DMU28	0.959	0.901	0.907	0.905	0.913	Decrease
29	DMU29	0.81	0.719	0.984	0.888	0.874	Decrease
30	DMU30	1	1	0.074	1	0.974	Decrease

Question 4: Is it possible to evaluate and compare students' performance using the Malmquist Index and Data Envelopment Analysis?

In this study, data envelopment analysis method and the Malmquist Index were used instead of using statistical methods, t-student test, etc., which reduced the computational volume and the execution of the model using Gams software, facilitated data analysis The final results, according to Table 7 in this study show the feasibility of using data envelopment analysis to compare the performance of students and classes and the Malmquist productivity evaluation Index to determine productivity growth and performance improvement over specified time periods, It can be used as a useful management tool for schools. All tangible and intangible factors are included and consider all factors, especially those that are intangible, according to their usability. For example, if only performance calculations were considered for students (tangible factors), in Class A, EMA = 23.3% and EAP = 36.6% were considered, as intangible factors were also considered. Class A productivity evaluation with the Malmquist Index was 63.3%, which can be attributed to increased student math literacy and Class A math progress.

Table 7. Class Evaluation Ranking

CLASS	Averag E(MA)	Averag E(AP)	Average(FGLR)	Rank(AP)	PIS
class A	23.30%	36.60%	63.30%	2	↑
class B	20%	33.30%	53.30%	3	↑
class C	13.30%	46.60%	80%	1	↑
class D	33.30%	20%	20%	4	↓

4. DISCUSSION and CONCLUSION

4.1. Conclusion

This paper deals with the integrated approach of data envelopment analysis method and the Malmquist Productivity Index to evaluate the performance of 9th grade high school students in different time intervals to enhance mathematical literacy. In this paper, the students' relative performance has been evaluated using data envelopment analysis and the Malmquist Index, and the rate of productivity and performance improvement has been determined in two different time intervals. According to the results, mathematics education can be successful when students have the ability to solve everyday real-world challenges based on mathematical facts, methods and concepts (Sumirattana et al., 2017). In solving a problem, the process of applying involves using mathematical knowledge directly. It is a process done in the mathematical world. Therefore, students who have done well in the world of mathematics do not have good results in the real world. Accordingly, students should be familiar with and skilled in mathematical processes such as problem-solving and applying problem-solving strategies and modelling (The national curriculum in, 2014).

On the other hand, a process happening in the math world can improve students' math performance in making connections in the real world. Therefore, it can be concluded that the combination of the two methods of problem-solving and realistic mathematics is effective in promoting mathematical literacy, given that mathematical skills and knowledge defined within the content of the mathematical curriculum are not intended in the use of the word "literacy". Its objective is to evaluate that part of mathematical knowledge that is used in many fields diversely, thoughtfully and insightfully. Mathematical literacy is not limited to mathematical technology knowledge, facts and mathematical calculations (Del R ó, Sanz, & B úcari, 2019) . In fact, we mean a wide, continuous and multi-dimensional spectrum from concepts to very high degrees. One of the most important competencies that are implicitly understood by the concept of mathematical literacy is the ability to design, formulate and solve internal and external mathematical problems in different domains (Alvarez, 2018). Some of the productivity indices in this research are easy understanding, access to information, easy calculation, and access to information to calculate such simple indicators and, if used in conjunction with the total productivity Index, are good tools for identifying weaknesses in the desired field.

The Malmquist Index and the GAMS can be used as a proper model to assess and rank the students in education system by resolving the deficiencies and limitations and by better identification of inputs and outputs since it is able to have a relatively fair evaluation of one way analyzes by examining the inputs and outputs simultaneously.

4.2. Discussion

One of the limitations of the Malmquist productivity indices is that they are very misleading if used alone and can be relatively difficult to obtain data needed for comparative purposes, so it is better to calculate them along with other models of data envelopment analysis. Despite the benefits of this Index to its limitations, it is recommended to use it when there are multiple inputs and outputs.

Because this Index provides a quantitative way to link everything from quality to process timing and dozens of other important performance indicators to profitability, it can be effective in the productivity evaluation of industry, agriculture and educational institutions. GAMS Software can be introduced as a good instrument for evaluating students' grades with different mathematical models in several time frames (Kalvelagen, n.d.), (Pintér, 2007).

ORCID

Niusha Mostoli https://orcid.org/0000-0001-6639-6878

Mohsen Rostamy https://orcid.org/0000-0001-6105-7674

5. REFERENCES

Adams, R., & Wu, M. (2000). *Pisa 2000 Technical (OECD Organization for Economic Co-operation and Developmant)*. Paris: Publications 2 rue André-Pascal, 75775 Cedex 16, France.

Alimohammadlou, M., & Mohammadi, S. (2016). Evaluating the Productivity Using Malmquist Index Based on Double Frontiers Data. Procedia - *Social and Behavioral Sciences*, 230(May), 58–66. https://doi.org/10.1016/j.sbspro.2016.09.008.

Alvarez, R. (2018) ."A Focus on Mathematical Literacy to Increase Student Understanding and Performance". (Master's dissertation) . New York, USA. Retrieved May5 , 2019 , from https://digitalcommons.brockport.edu/ehd_theses/1169.

Andersen, P., & Petersen, N. C. (1993). A Procedure for Ranking Efficient Units in Data Envelopment Analysis process (DEA). *Management Science*, *39*(10), 1261-1264. https://doi.org/10.2307/2632964.

Basic, T. H. E., & Model, C. C. R. (2005). The Basic CCR Model. *In Data Envelopment Analysis* (pp. 21–39). Boston: Kluwer Academic Publishers. https://doi.org/10.1007/0-306-47541-3_2.

Breen, S., Cleary, J., & Shea, A. O. (2009). An investigation of the Mathematical literacy of first year third-level students in the Republic of Ireland. *International Journal of Mathematical Education, 40*(2) , 229-246. https://doi.org/10.1080/00207390802566915.

Burnett, N. (2005). literacy for life. *Education for All* . France: Printed by Graphoprint, Paris ISBN 92-3-104008-1©UNESCO .https://unesdoc.unesco.org/ark:/48223/pf0000141639.

Chen, C., & Lin, M. (2006). Using DEA to Evaluate R & D Performance in the Integrated Semiconductor Firms - Case Study of Taiwan, *International Journal of The Computer the Internet and Management, 14*(3), 50–59.

Del Río, L. S., Sanz, C. V., & Búcari, N. D. (2019). Incidence of a hypermedia output material on the teaching and learning of mathematics. *Journal of New Approaches in Educational Research, 8*(1), 50–57. https://doi.org/10.7821/naer.2019.1.334.

Diewert, W. E., & Fox, K. J. (2010). Malmquist and Törnqvist productivity indexes: Returns to scale and technical progress with imperfect competition. *Journal of Economics/ Zeitschrift Fur Nationalokonomie, 101*(1), 73–95. https://doi.org/10.1007/s00712-010-0137-0.

Douglas, W., Caves ,E., Laurits, R., & Christensen, W. E. D. (2012). *The Economic Theory of Index Numbers and the Measurement of Input , Output , and Productivity* Author (s): Douglas W . Caves , Laurits R . Christensen , W . Erwin Diewert Reviewed work (s):

Published by : *50*(6), 1393–1414.

Esther, M., Pérez, M., Duque, A. P. G., & García, L. C. F. (2018). Game-Based Learning : Increasing the Logical-Mathematical , Naturalistic , and Linguistic Learning Levels of Primary School Students, *7*(1), 31–39. https://doi.org/10.7821/naer.2018.1.248.

Fare, R., Grosskopf, S., Lindgren, B., Roos, P. (1992). A Non-Parametric Malmquist Approach. *The Journal of Productivity Analysis*, 3(1), 85–101. Retrieved from https://link.springer.com/content/pdf/10.1007%2FBF00158770.pdf.

Färe, R., Grosskopf, S., & Roos, P. (1995). Productivity and quality changes in Swedish pharmacies. *International Journal of Production Economics*, 39(1–2), 137–144. https://doi.org/10.1016/0925-5273(94)00063-G.

Freudenthal, H. (1973). Mathematics as an educational task, Reidel. Van den Heuvel-Panhuizen, M. (2000). *Mathematics education in theNetherlands: A guided tour. Freudenthal Journal*, Utrecht University.

Gatabi, A. R., Stacey, K., & Gooya, Z. (2012). Investigating grade nine textbook problems for characteristics related to mathematical. *literacy Math Ed Res Journal*, (2012) 24, 403–421. https://doi.org/10.1007/s13394-012-0052-5.

Gravemeijer, K. P. E. (1994). Developing realistic mathematics education. *Journal of Curriculum Studies* ,(1994) 11, 24-34. https://research.tue.nl/en/publications/developing-realistic-mathematics-education.

Gravemeijer, K., Terwel, J.(2000). Hans Freudenthal: a mathematician on didactics and curriculum theory. *Journal of Curriculum Studies2000*, VOL. 32, NO. 6, 777 ± 796.

Hosseinzadeh Lotfi, F., Jahanshahloo, G. R., Khodabakhshi, M., Rostamy-Malkhlifeh, M., Moghaddas, Z., & Vaez-Ghasemi, M. (2013). A Review of Ranking Models in Data Envelopment Analysis. *Journal of Applied Mathematics*, 2013 NO. (14), 1–20. https://doi.org/10.1155/2013/492421.

Kalvelagen, E. (2002). Data and Software Interoperability with Gams :A UserPerspectiveModelingLanguages, *Solving Multi-Objective Models with GAMS Journal*, http://citeseerx.ist.psu.edu/ doi=10.1.1.201.332&rep=rep1&type=pdf

Koçak, D., Türe, H., Atan, M. (2019). Efficiency Measurement with Network DEA : An Application to Sustainable Development Goals. *International Journal of Assessment Tools in Educatio*, 6(3), 415–435.

Koyuncu, İ., Guzeller, C. O., & Akyuz, D. (2016). The development of a self-efficacy scale for mathematical modeling competencies. *International Journal of Assessment Tools in Education*, *4*(1), 19–35. https://doi.org/10.21449/ijate.256552

Lange, J. De. (1987). *Mathematics, Insight and Meaning: Teaching, Learning and Testing of Mathematics for the Life and Social Sciences*. Freudenthal Institute. Retrieved from https://books.google.com/books?id=wee5pwAACAAJ.

Maudos, J., & Pastor, J. M. (1998). *Human capital in OECD countries* : Technical change , efficincy and Joaquín Maudos , José Manuel Pastor and Lorenzo Serrano.

OECD. (2002). *Reading for Change: Performance and Engagement across Countries*. OECD. https://doi.org/10.1787/9789264099289-en.

Pintér, J. D. (2007). Nonlinear optimization with GAMS /LGO. *Journal of Global Optimization*, *38*(1), 79–101. https://doi.org/10.1007/s10898-006-9084-2.

Polya, G. (1962). Mathematical discovery: On understanding, learning and teaching problem solving. *New York: Wiley*, (vol. 1).

Programme for International Student Assessment., & Organisation for Economic Co-operation and Development. (2009). *Learning mathematics for life : a perspective from PISA*. OECD.

Rosenthal, R. E. (2007). A GAMS Tutorial. *GAMS - A User's Guide*, 5–26.

Sari, Y. M., & Valentino, E. (2016). An Analysis of Students Error In Solving PISA 2012 And

Its Scaffolding, *1*(2), 90–98.

Sinuany-Stern, Z., Mehrez, A., & Barboy, A. (1994). Academic departments efficiency via DEA. *Computers & Operations Research*, *21*(5), 543–556. https://doi.org/10.1016/0305-0548(94)90103-1.

Stacey, K., Almuna, F., Caraballo, R. M., Lupia, L., Park, K. M., Perl, H., & Rafiepour, A. (2015). PISA' s Influence on Thought and Action in Mathematics Education. https://doi.org/10.1007/978-3-319-10121-7.

Sumirattana, S., Makanong, A., & Thipkong, S. (2017). Using realistic mathematics education and the DAPIC problem-solving process to enhance secondary school students' mathematical literacy. *Kasetsart Journal of Social Sciences*, *38*(3), 307–315. https://doi.org/10.1016/j.kjss.2016.06.001.

The national curriculum in. (2014, December), https://www.gov.uk/dfe/nationalcurriculum .

Thomson, S., Hillman, K., & Bortoli, L. De. (2013). A Teacher's Guide to PISA Mathematical Literacy. *OECD Programme for International Student Assessment (PISA Australia)*. Retrieved from https://research.acer.edu.au/ozpisa/12.

Tohidi, G., & Razavyan, S. (2013). A circular global profit Malmquist productivity index in data envelopment analysis. *Applied Mathematical Modelling*, *37*(1–2), 216–227. https://doi.org/10.1016/j.apm.2012.02.026.

Wang, Y. M., & Lan, Y. X. (2011). Measuring Malmquist productivity index: A new approach based on double frontiers data envelopment analysis. *Mathematical and ComputerModelling*, *54*(11-12), 2760-2771. https://doi.org/10.1016/j.mcm.2011.06.064

Wei, Z., & Hao, R. (2011). The Role of Human Capital in China's Total Factor Productivity Growth . *The Developing Economies*, 49(1), 1-35.

6. APPENDIX

Score INPUT & OUTPUT Any Qestion

An example of the problem based on realistic mathematics education and PISA problem-solving (2015):

A pizza seller sells two types of pizza with the same thickness in a small size with a diameter of 30 cm, and a price of 30$, in a large size with a diameter of 40 cm, at a price of 40 $, Which pizza is worth buying?

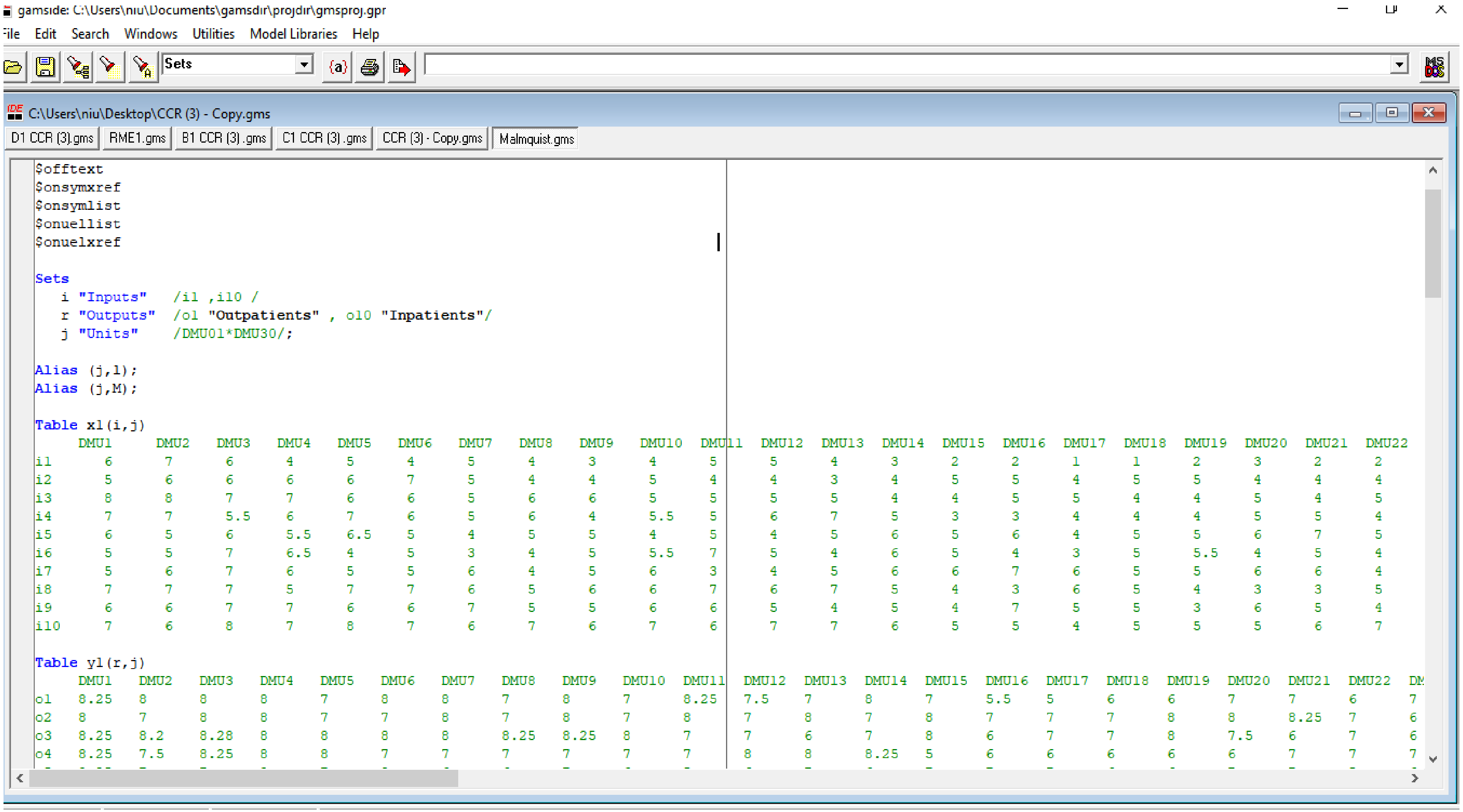
DAPIC Math Problem-solving

1. **Define**: definition of the problem with the student experiences, What do I want to do (Pizza shape, definition of the circle) 0.25
2. **Assess**: Separation of keywords in the problem, assuming data as hypotheses. (Definition of the area and perimeter of the circle, smaller and larger) 0.25
3. **Plan**: Guessing the operation, designing and determining the program. (Which approach is appropriate) 0.25
4. **Implement**: implementation of plans and guesses, make changes as needed. (Find the perimeter and area and the appropriate ratio) 0.25
5. **Communicate**: assessing and analyzing the result, reviewing the problem back. (Review and analysis of the solution) 0.25

Realistic Mathematics Education (RME)

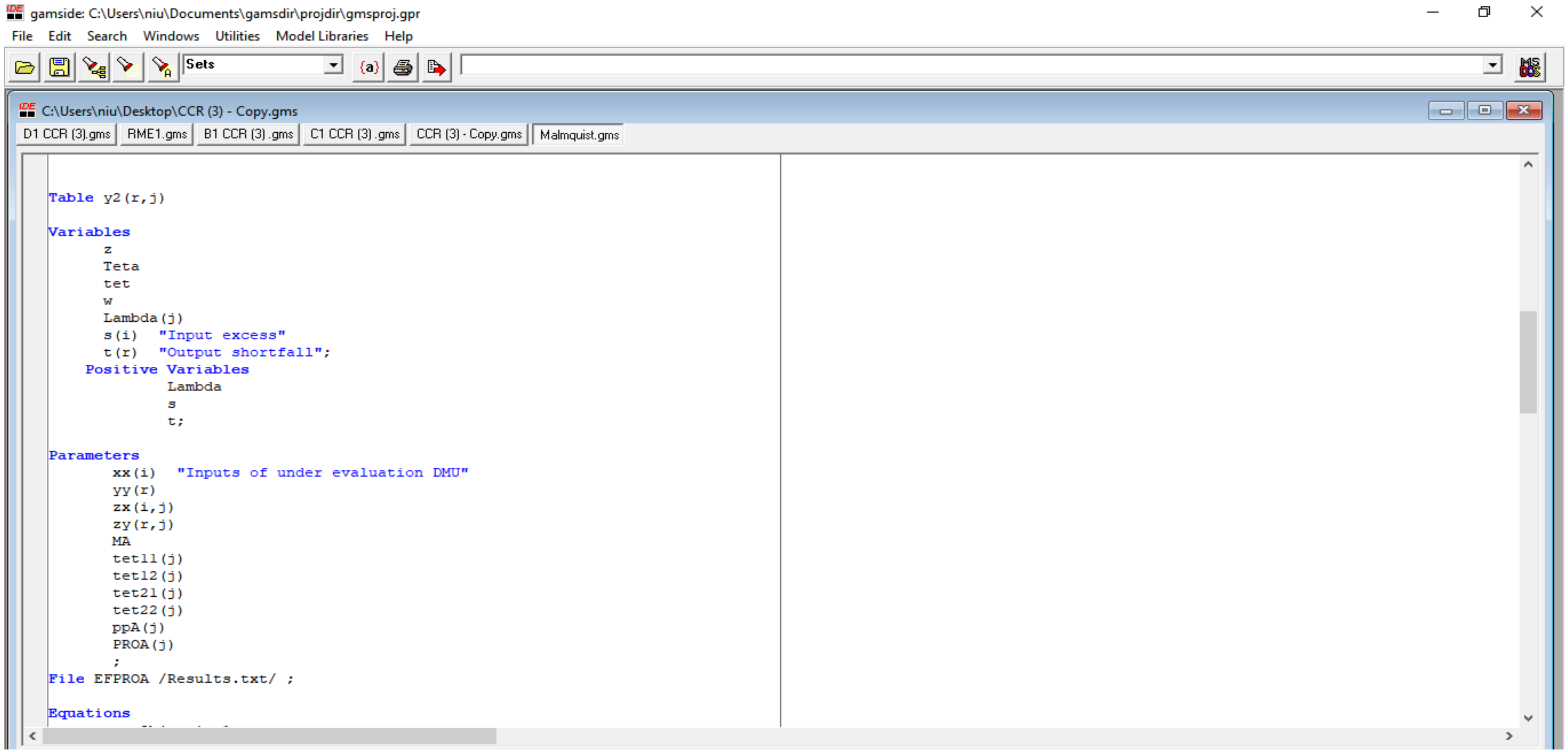
1. **Guided reinvention**: a hypothesis to create knowledge and innovation (innovation and communication with the real world) 0.25
2. **Didactical phenomenology**: Creating a comprehensive description of an experienced daily and real life phenomenon (the real experience of buying pizza) 0.25
3. **Self-developed model**: The model is developed by the student himself from an informal to formal math. (Communication of Reality with Math) 0.25

GAMS Software with the Malmquist Index

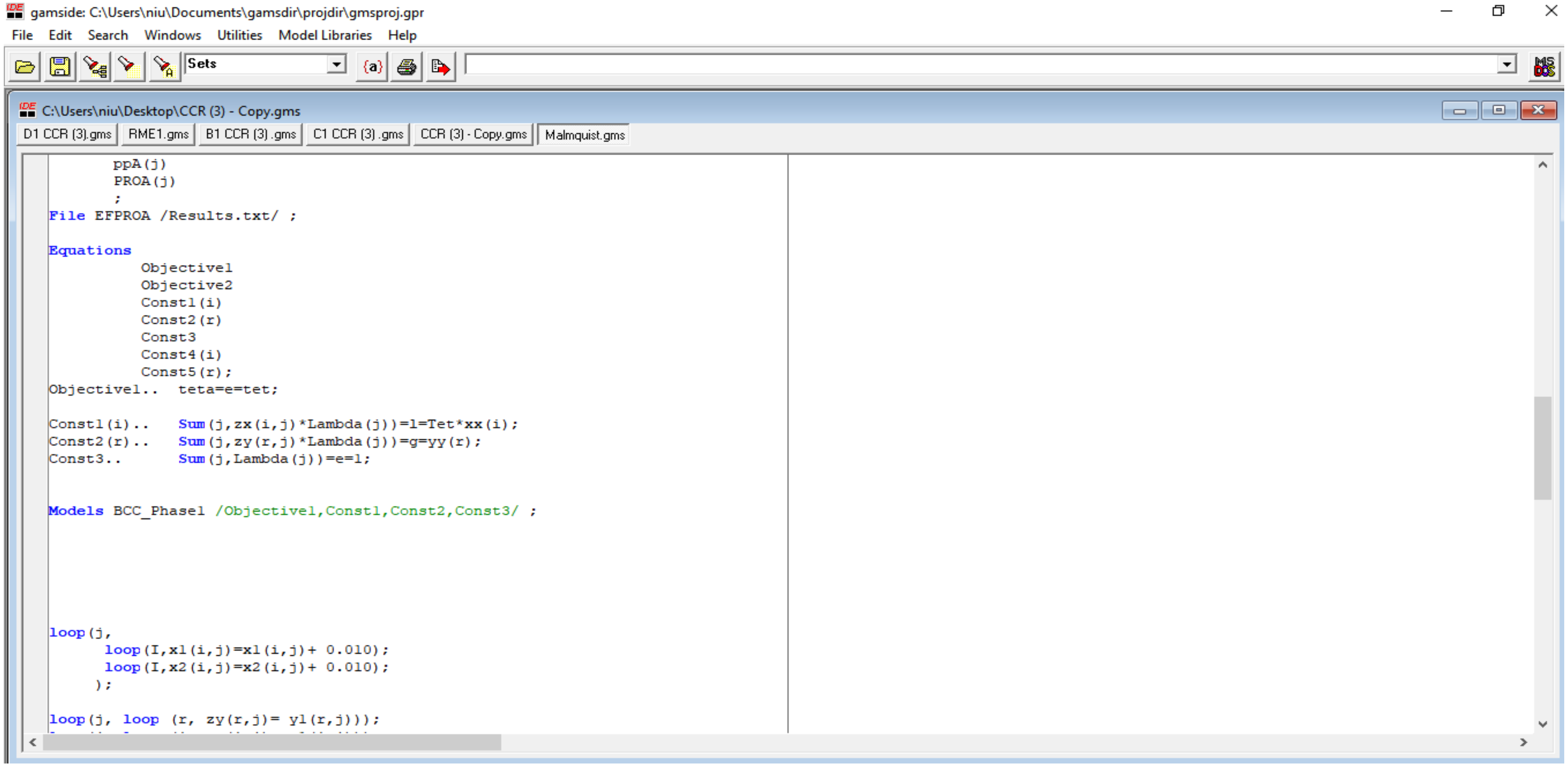


gamside: C:\Users\niu\Documents\gamsdir\projdir\gmsproj.gpr
File Edit Search Windows Utilities Model Libraries Help
Sets
C:\Users\niu\Desktop\CCR (3) - Copy.gms
D1 CCR (3).gms
RME1.gms
B1 CCR (3) .gms
C1 CCR (3) .gms
CCR (3) - Copy.gms
Malmquist.gms

```
$offtext
$onsymxref
$onsymlist
$onuellist
$onuelxref

Sets
   i "Inputs"   /i1 ,i10 /
   r "Outputs"  /o1 "Outpatients" , o10 "Inpatients"/
   j "Units"    /DMU01*DMU30/;

Alias (j,l);
Alias (j,M);

Table x1(i,j)
     DMU1      DMU2   DMU3   DMU4   DMU5   DMU6   DMU7   DMU8   DMU9   DMU10  DMU11  DMU12  DMU13  DMU14  DMU15  DMU16  DMU17  DMU18  DMU19  DMU20  DMU21  DMU22
i1      6        7      6      4      5      4      5      4      3      4      5      5      4      3      2      2      1      1      2      3      2      2
i2      5        6      6      6      6      7      5      4      4      5      4      4      3      4      5      5      4      5      5      4      4      4
i3      8        8      7      7      6      6      5      6      6      5      5      5      5      4      4      5      5      4      4      5      4      5
i4      7        7      5.5    6      7      6      5      6      4      5.5    5      6      7      5      3      3      4      4      4      5      5      4
i5      6        5      6      5.5    6.5    5      4      5      5      4      5      4      5      6      5      6      4      5      5      6      7      5
i6      5        5      7      6.5    4      5      3      4      5      5.5    7      5      4      6      5      4      3      5      5.5    4      5      4
i7      5        6      7      6      5      5      6      4      5      6      3      4      5      6      6      7      6      5      5      6      6      4
i8      7        7      7      5      7      7      6      5      6      6      7      6      7      5      4      3      6      5      4      3      3      5
i9      6        6      7      7      6      6      7      5      5      6      6      5      4      5      4      7      5      5      3      6      5      4
i10     7        6      8      7      8      7      6      7      6      7      6      7      7      6      5      5      4      5      5      5      6      7

Table y1(r,j)
     DMU1   DMU2   DMU3   DMU4   DMU5   DMU6   DMU7   DMU8   DMU9   DMU10  DMU11  DMU12  DMU13  DMU14  DMU15  DMU16  DMU17  DMU18  DMU19  DMU20  DMU21  DMU22  DM
o1   8.25   8      8      8      7      8      8      7      8      7      8.25   7.5    7      8      7      5.5    5      6      6      7      7      6      7
o2   8      7      8      8      7      7      8      7      8      7      8      7      8      7      8      7      7      7      8      8      8.25   7      6
o3   8.25   8.2    8.28   8      8      8      8      8.25   8.25   8      7      7      6      7      8      6      7      7      8      7.5    6      7      6
o4   8.25   7.5    8.25   8      8      7      7      7      7      7      7      8      8      8.25   5      6      6      6      6      6      7      7      7
```



```
Table y2(r,j)

Variables
      z
      Teta
      tet
      w
      Lambda(j)
      s(i)   "Input excess"
      t(r)   "Output shortfall";
    Positive Variables
             Lambda
             s
             t;

Parameters
       xx(i)  "Inputs of under evaluation DMU"
       yy(r)
       zx(i,j)
       zy(r,j)
       MA
       tet11(j)
       tet12(j)
       tet21(j)
       tet22(j)
       ppA(j)
       PROA(j)
       ;
File EFPROA /Results.txt/ ;

Equations
```



```
       ppA(j)
       PROA(j)
       ;
File EFPROA /Results.txt/ ;

Equations
          Objective1
          Objective2
          Const1(i)
          Const2(r)
          Const3
          Const4(i)
          Const5(r);
Objective1..  teta=e=tet;

Const1(i)..   Sum(j,zx(i,j)*Lambda(j))=l=Tet*xx(i);
Const2(r)..   Sum(j,zy(r,j)*Lambda(j))=g=yy(r);
Const3..      Sum(j,Lambda(j))=e=1;


Models BCC_Phase1 /Objective1,Const1,Const2,Const3/ ;






loop(j,
      loop(I,x1(i,j)=x1(i,j)+ 0.010);
      loop(I,x2(i,j)=x2(i,j)+ 0.010);
     );

loop(j, loop (r, zy(r,j)= y1(r,j)));
```

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 656–669
https://dx.doi.org/10.21449/ijate.635675

Published at http://www.ijate.net http://dergipark.org.tr **_Research Article_**

Developing an Item Bank for Progress Tests and Application of Computerized Adaptive Testing by Simulation in Medical Education

Ayşen Melek Aytuğ Koşan[1,*] **Nizamettin Koç**[2], **Atilla Halil Elhan**[3], **Derya Öztuna**[3]

[1]Çanakkale Onsekizmart University; School of Medicine, Medical Education and Informatics Department, Çanakkale, Turkey
[2]Ankara University School of Education, Measurement and Evaluation Department, Ankara, Turkey
[3]Ankara University School of Medicine, Biostatistics Department, Ankara, Turkey

ARTICLE HISTORY

Received: 21 October 2019
Revised: 26 November 2019
Accepted: 17 December 2019

KEYWORDS

Computer Adaptive Test,
Rasch Models,
Progress Test,
Medical Education,
Item Bank Development

Abstract: Progress Test (PT) is a form of assessment that simultaneously measures ability levels of all students in a certain educational program and their progress over time by providing them with same questions and repeating the process at regular intervals with parallel tests. Our objective was to generate an item bank for the PT and to examine the possible fit of CAT for PT application. This study is a descriptive study. 1206 medical students participated. During the analysis of the psychometric properties of PT item bank, "the Rasch model for dichotomous items was used". Several CAT simulations were performed by applying various stopping rules of different standard errors. CAT simulation estimates were compared with the estimates generated from the original calibration of the Rasch model where all items were included. After Rasch analysis, a unidimensional PT item bank consisting of 103 items was obtained. The item bank reliability was calculated as 0.77 with Person Separation Index (PSI) and Kuder-Richardson Formula 20 (KR-20). A high correlation between θ estimations obtained from paper-and-pencil (θRM) and CAT applications (θ_{CAT}) was detected for simulation conditions ([N(0,1)] and [N(0,3)]) at the end of our analysis. In CAT, estimation can be made with an average of 14 questions (reduced 86,4%) and 17 questions (reduced 83,4%) [for N(0,1) and [N(0,3) respectively] with reliability of 0,75. This study reveals that it is possible to develop an appropriate item bank for the PT, and the difficulty of administering large number of items in PT can be scaled down by incorporating CAT application.

1. INTRODUCTION

A Progress Test (PT) is a type of assessment that simultaneously measures the ability levels of all students in a certain educational program and their progress over time during that program by providing them with the same questions and repeating the process at regular intervals with parallel tests (Freeman, 2010). Following a blueprint geared toward the cognitive learning objectives anticipated at the end of the curriculum, a question sample that is representative of

CONTACT: Ayşen Melek AYTUĞ KOŞAN ✉ aysenay1@yahoo.com 🖃 Canakkale Onsekizmart University; School of Medicine, Medical Education and Informatics Department, Canakkale, TURKEY

ISSN-e: 2148-7456 /© IJATE 2019

all the disciplines and content areas is used in the PT. Due to its contributions to education, PT is used in medical education worldwide, however it has numerous drawbacks as it demands lots of manual effort from the human resources based on the time needed for its preparation, the implementation itself, and evaluation of the results. It also consists of numerous questions, making it a less attractive method for students as it results in their exhaustion (Wrigley. 2012). Implementing PT with the Computerized Adaptive Testing (CAT) method would cut down on the time and human effort needed in the evaluation of medical knowledge as it helps limit the number of items. This also reduces the bulkiness of the process and in turn the negative feelings associated with it.

Purpose of the study; this study aims to develop an item bank for the PT used by the Ankara University School of Medicine (AUSM) and to investigate the capability of CAT for evaluating the medical knowledge in AUSM students

2. METHOD

Sample/Working Group/Participants

This study is a descriptive study. PT is a component of the evaluation system in AUSM. Participants volunteered for PT and received bonus points on their final examinations. The data used for this study was obtained from the results of a PT taken in the 2010-2011 academic year. One thousand two hundred six medical students participated in the PT in grades 1-5 at the AUSM (89.7% of the total) in 2011.

Data Collection Instruments/Data Collection Methods/Data Collection Techniques

The study was divided into two phases: (i) the development of an item bank using a dichotomous Rasch model; (ii) a simulation study to investigate the performance of CAT application with stopping rules of various standard errors or reliability values; and examination of agreement between ability estimates derived from simulated CAT (θ_{CAT}) application and ability estimates from the Rasch model (θ_{RM}) derived from all the items included in the original calibration.

2.1.1. *Development of item bank*

The PT consisted of “200 multiple-choice questions in single best answer format”. Items within each test were classified and matched to the blueprint. The most important component of a CAT is an Item Response Theory (IRT)-based unidimensional and calibrated item bank (Abberger, 2013; Wright & Bell, 1984). To develop an item bank, all 200 items in the chosen PT were considered as candidate items. The first stage of the study, an IRT model, was constructed while examination of the psychometric characteristics of the item bank.

2.1.2. *IRT model selection*

One of the main models of IRT, the Rasch model, produces two different estimates; “the latent trait person estimates” which are independent of the population distribution, and the “item difficulty estimates” which are independent of the person’s ability (Andrich, 1988). Examination of the psychometric properties of the item bank and estimation of item parameters, the Rasch model for dichotomous items was performed using the RUMM 2020 software (Andrich, 2003).

To determine the observed data to fit in the Rasch model, numerous statistical processes are used. In this respect, the Rasch analysis included the following steps in this study:

- Distractor analysis
- Unidimensionality and local independence
- Item-model fit
- Invariance of item parameters

- Differential item functioning (DIF)
- Internal consistency reliability of the item bank (Person Separation Index-[PSI] and Kuder-Richardson 20 [KR-20])

2.1.3. *Distractor Analysis*

A Distractor Analysis examines the distractor curves' trend consistent with the Item Characteristics Curve (ICC). As the ability level of the student increases, the correct distractor should follow the general shape of the ICC. Students with higher ability level will likely choose the correct distractor, while the probability value of the other distractors will decrease. Prior to item calibration, item analysis was performed to identify potential item problems by Rasch for dichotomous data (Andrich, 2003).

2.1.4. *Unidimensionality and local independence*

Items to be entered into the item bank were also required to meet unidimensionality and local independence assumptions. Unidimensionality means that all items the test is composed of measure only a single construct.

In this study Principle Component Analysis (PCA) of residuals obtained from the Rasch model was used to examine the unidimensionality assumption. In PCA analysis, if there is no meaningful pattern in the residuals, then it is concluded that the unidimensionality assumption is met. PCA of residuals comprised an item residual correlation matrix. Through this matrix, the correlation between the items and the first residual factor are examined to identify two subsets of items (the positively and negatively correlated items). Difference between the each person estimate obtained from these positively and negatively correlated item sets is compared by independent t-tests. To meet the unidimensionality assumption, the percentage of tests outside the range ±1.96 should not to exceed 5% of total number of tests (Elhan, 2010; Pallant & Tennant, 2007; Tennant & Pallant, 2006).

2.1.5. *Test of fit to the model*

"Rasch item fit" statistics showed how accurately the test data fit the Rasch measurement model (Linacre, 2000). Overall quality of fit for Rasch models was measured regarding the following:

- Overall item fit statistic
- Overall person fit statistics
- Item-trait interaction

Overall item and overall person fit statistics transformed to a z-score. If the items and person data meet the model expectation, it is anticipate that the mean will be approximately zero and the standard deviation will be one.

Other fit statistic which is applies in this study is an item-trait interaction statistic. This statistic is reported as a chi-square and showed the characteristic of invariance across the trait. A non-significant chi-square indicates that the hierarchical ordering of the items do not vary across the trait, denoting the requirement of invariance is met.

Further to these overall summary fit statistics, individual item fit statistics and person fit statistics were applied by using residuals and chi-square statistics. In the individual item fit statistics and person fit statistics that are based on the standardized residuals, the computed residual value of "z" should range between ±2.5, indicating a satisfactory fit to the model. Consequently, the tests of the individual item/person fit were also conducted based on chi-squares. For a given item/person, several chi-squares are computed, and then these chi-square values are totaled to give the overall chi-square for the item. If the p value calculated from the overall chi-square is less than 0.05 (or Bonferroni adjusted value), then the item is considered unfit for the model (Öztuna, 2008; Pallant & Tennant, 2007; Tennant & Conaghan, 2007). Bonferroni corrections were implemented to fit statistics (Bland & Altman,1995).

2.1.6. *Differential item functioning*

To be entered into the item bank, items were required to be free of differential item functioning (DIF). The model fit is affected by item bias. DIF appears when different groups score differently on a specific item, given the same location value of the latent trait (Andrich & Hagquist, 2012; Hambleton, 1991; Teresi, 2000). In this study, a variance-based statistical analysis was performed to test DIF and artificial DIF by grades by using RUMM 2020 software (Andrich, 2003).

2.1.7. *Reliability and content validity of item bank*

Reliability was studied with the Person Separation Index (PSI) and Kuder-Richardson Formula 20 (KR-20). PSI indicated whether the test discriminates students into groups according to their ability, and a PSI of 0.7 or more evidence a fit with the Rasch model KR-20 ranged between 0 and 1, where the value of 1 indicated perfect reproducibility of person placements (Fisher, 1992; Nunnally & Bernstein, 1994; Tavakol & Dennick, 2012). Experts from different medical specialties and measurement and evaluation experts examined the content validity of the item bank.

2.1.8. *Simulation study to investigate the performance of CAT application and agreement between Rasch and simulated CAT derived estimations Computerized Adaptive Testing (CAT)*

After developing the calibrated item bank, the next stage of simulated CAT application was carried out. In CAT application, a set of questions was administered to each student according to their ability level by using a computer package program. For this purpose, the questions in the item bank with the median difficulty level were administered, and the program estimated the students' ability level (θ_{CAT}) and its standard error. After this estimation, the next most appropriate item was selected that maximized the information for θ estimate, and then the program re-estimated the students' ability level (θ_{CAT}) and its standard error. The CAT application program selected the questions for each student, according to his or her individual performance during the test. If the predefined stopping rule was fulfilled, the assessment was finished; if it was not fulfilled, the standard error of the given item administered and the ability level were re-estimated until the stopping rule was met (Bjorner, 2007; Wainer, 2001).

2.1.9. *Simulated CAT applications*

In this study item parameters obtained from the Rasch analysis were used to derive responses of 1000 students/simulee showing two different normal distributions with N(0:1) (mean=0, standard deviation=1) and N(0:3) (0 mean=0, standard deviation=3) by the RUMMss simulation software. These data were simulated to meet Rasch model expectations (Marais & Andrich, 2007). Students' responses generated by the simulation program were used to estimate student ability level using all the items (θ_{RM}), while student ability levels (θ_{CAT}) were estimated by CAT application using the SmartCAT module (Öztuna, 2008; 2012).

Selection of the first question: The question with the median difficulty level in the item bank

- Ability level (θ) estimation: Expectation a Posteriori (EAP) (Wang &Vispoel, 1998)
- Item selection: Maximum Likelihood Weight Information
- Stopping rule: Different standard errors levels (0.50, 0.40 and 0.30)

Estimations from the simulated CAT application (θ_{CAT}) were compared with the ability levels obtained from the Rasch analysis based on all items (θ_{RM}) in the item bank. In this procedure, Spearman correlation coefficient, Bland-Altman limits of agreement (Bland & Altman, 1986; 1999), and Interclass Correlation Coefficient (ICC) statistics were used (Shrout & Fleiss, 1979).

3. RESULTS

A total of 1206 students of AUSM answered 200 items of PT. Distractor analysis was conducted on these 200 items and because of the problems such as “too obvious correct answer” and “discordance pattern of correct answer with ICC”, 33 items were discarded from the item bank (Figure 1). Analyses were performed on the remaining 167 items.

Table 1. Fit of "general medical knowledge” item bank to Rasch model (after rescoring)

Item No	B	SE	Individual Item Fit Residual	X^2	df	p
1 (i1)	0.450	0.059	2.712	4.047	9	0.908
2 (i2)	0.666	0.059	1.835	5.951	9	0.745
3 (i)	0.161	0.060	-1.078	9.683	9	0.377
4 (i5)	-0.124	0.062	1.751	10.036	9	0.348
5 (i7)	-0.286	0.064	-0.627	5.949	9	0.745
6 (m8)	-0.146	0.063	0.531	3.161	9	0.958
7 (i9)	1.242	0.062	0.439	7.917	9	0.543
8 (i10)	0.988	0.060	1.027	14.087	9	0.119
9 (i12)	0.883	0.059	1.377	7.983	9	0.536
10 (i13)	1.047	0.060	1.967	10.128	9	0.340
11 (m15)	-0.271	0.064	-1.685	12.958	9	0.165
12 (m19)	1.079	0.060	-0.913	10.384	9	0.320
13 (i21)	-2.769	0.154	-0.364	11.385	9	0.250
14 (i22)	0.809	0.059	1.266	4.721	9	0.858
15 (i24)	-0.349	0.065	1.223	15.112	9	0.088
16 (i25)	-0.547	0.068	1.118	16.137	9	0.064
17 (i26)	-1.334	0.085	-0.578	7.916	9	0.543
18 (i30)	-0.673	0.070	0.921	14.898	9	0.094
19 (i31)	0.201	0.060	2.894	10.344	9	0.323
20 (i33)	0.039	0.061	-0.207	10.740	9	0.294
21 (i35)	0.249	0.060	3.432	20.162	9	0.017
22 (i36)	0.944	0.060	2.187	10.416	9	0.318
23 (m37)	-0.281	0.064	2.182	10.786	9	0.291
24 (i39)	0.471	0.059	-1.055	6.963	9	0.641
25 (i41)	-0.221	0.063	0.363	9.704	9	0.375
26 (i43)	-0.747	0.071	0.290	5.509	9	0.788
27 (i44)	0.395	0.059	2.411	13.970	9	0.123
28 (i45)	-1.465	0.089	0.094	9.199	9	0.419
29 (i46)	0.569	0.059	1.739	7.000	9	0.637
30 (i48)	-0.194	0.063	0.568	8.184	9	0.516
31 (i50)	-0.686	0.070	-0.082	6.296	9	0.710
32 (i52)	0.015	0.061	1.565	11.500	9	0.243
33 (i54)	0.450	0.059	0.222	17.966	9	0.036
34 (i55)	-1.791	0.101	-0.563	11.493	9	0.243
35 (i56)	-0.389	0.065	-0.717	8.558	9	0.479
36 (i58)	0.050	0.061	2.614	16.004	9	0.067
37 (i59)	-0.049	0.062	0.130	6.533	9	0.686
38 (i60)	-0.997	0.076	-1.006	8.024	9	0.532
39 (i61)	-1.450	0.089	-0.344	8.329	9	0.501
40 (i62)	-0.279	0.064	-1.013	12.447	9	0.189
41 (i63)	-1.157	0.080	0.307	16.995	9	0.049

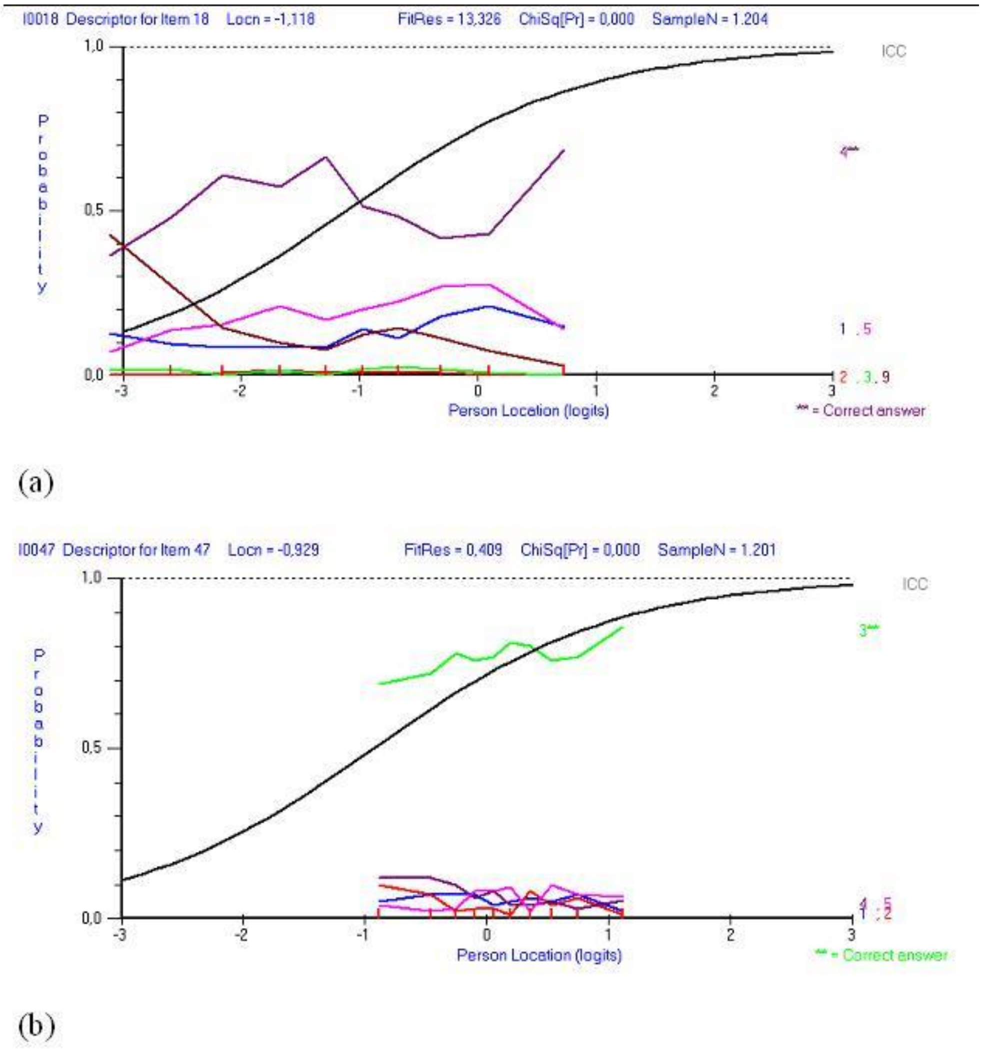
Table 1. Continues

42 (i65)	-2.064	0.113	0.238	9.990	9	0.351
43 (i69)	-0.405	0.066	0.030	7.078	9	0.629
44 (i70)	0.257	0.060	-0.040	11.881	9	0.220
45 (i73)	0.204	0.060	2.389	6.662	9	0.672
46 (i76)	1.123	0.061	1.872	10.796	9	0.290
47 (i79)	-0.311	0.064	1.186	5.011	9	0.833
48 (i87)	-0.274	0.064	-0.481	9.212	9	0.418
49 (i88)	-0.940	0.075	-0.706	9.410	9	0.400
50 (i89)	1.415	0.063	0.721	12.853	9	0.169
51 (m92)	-0.150	0.063	-0.029	10.393	9	0.320
52 (m102)	0.167	0.060	-1.330	10.783	9	0.291
53 (i103)	0.725	0.059	-0.572	14.471	9	0.107
54 (i104)	0.283	0.060	-1.840	12.909	9	0.167
55 (i105)	-1.646	0.096	-0.916	18.944	9	0.026
56 (i106)	1.431	0.064	0.055	4.844	9	0.848
57 (i113)	-0.228	0.063	-0.738	11.141	9	0.266
58 (i114)	0.801	0.059	0.885	5.515	9	0.787
59 (i115)	0.167	0.060	-0.269	6.465	9	0.693
60 (i116)	1.131	0.061	1.534	13.360	9	0.147
61 (i117)	0.198	0.060	1.845	7.156	9	0.621
62 (i119)	0.502	0.059	-1.374	17.268	9	0.045
63 (i121)	0.358	0.059	0.590	5.691	9	0.770
64 (i122)	0.322	0.059	-1.923	18.530	9	0.030
65 (i123)	-1.127	0.080	-1.039	12.310	9	0.196
66 (i124)	-1.177	0.081	-1.308	16.593	9	0.055
67 (i125)	-0.296	0.064	1.033	9.038	9	0.434
68 (i128)	-0.269	0.064	-1.051	10.836	9	0.287
69 (i129)	0.616	0.059	2.739	12.926	9	0.166
70 (i130)	0.561	0.059	0.957	11.488	9	0.244
71 (i131)	-0.765	0.072	-1.099	9.396	9	0.402
72 (i133)	0.743	0.059	-1.062	10.213	9	0.334
73 (i139)	0.191	0.060	1.022	16.195	9	0.063
74 (i140)	-1.369	0.086	-0.807	9.810	9	0.366
75 (i141)	0.879	0.059	1.681	8.408	9	0.494
76 (i142)	1.127	0.061	1.700	7.514	9	0.584
77 (i145)	1.299	0.062	2.395	15.809	9	0.071
78 (i148)	0.687	0.059	0.771	4.412	9	0.882
79 (i151)	-0.695	0.070	-0.107	5.019	9	0.833
80 (m155)	-0.103	0.062	0.011	5.553	9	0.784
81 (i156)	-0.649	0.069	-0.254	14.803	9	0.096
82 (i157)	-0.116	0.062	-0.003	5.754	9	0.764
83 (i158)	0.139	0.060	0.490	10.426	9	0.317
84 (i160)	0.064	0.061	0.884	6.378	9	0.702
85 (i161)	0.718	0.059	2.828	18.111	9	0.034
86 (i165)	-0.278	0.064	0.167	7.022	9	0.635
87 (i166)	0.718	0.059	0.650	13.473	9	0.142
88 (m171)	0.180	0.060	0.138	11.755	9	0.227
89 (i172)	-0.672	0.070	0.355	8.026	9	0.532
90 (i177)	0.329	0.059	1.157	6.059	9	0.734

Table 1. Continues

91 (i178)	0.494	0.059	0.572	6.733	9	0.665
92 (i180)	0.209	0.060	2.800	11.833	9	0.223
93 (i181)	0.653	0.059	-0.988	6.108	9	0.729
94 (i182)	0.104	0.061	0.932	8.226	9	0.512
95 (m183)	0.376	0.059	-1.006	10.950	9	0.279
96 (i185)	1.341	0.063	0.137	14.669	9	0.100
97 (i188)	0.159	0.060	-0.433	10.147	9	0.339
98 (i191)	-0.068	0.062	0.280	2.790	9	0.972
99 (i192)	0.173	0.060	1.213	11.649	9	0.234
100 (i193)	-0.370	0.065	-0.896	14.136	9	0.118
101 (i194)	-1.185	0.081	0.159	6.196	9	0.720
102 (i195)	-0.888	0.074	-0.318	7.108	9	0.626
103 (i200)	0.696	0.059	-1.364	11.517	9	0.242

B: Item Difficulty, SE: Standard Error, df= Degrees of Freedom



(a)

(b)

Figure 1. Item analysis og item 18 (a) and 47 (b)

3.1. Development of item bank (internal construct validity)

In order to be entered into the item bank, items were required to satisfy Rasch model expectations, including being free of DIF and having unidimensionality and local independence. Sixty-four of 167 items were omitted as not fitting the Rasch model. For the remaining 103 items, p values from chi-square were less than Bonferroni adjustment fit level of 0,0005 (0.05/103=0,0005). In addition, the standard residual values were within the ±2.5 range. These statistics indicated that the items fit the Rasch model (Table 1).

When the overall fit statistics were tested, it was found that the overall mean item fit residual was 0.402 (SD 1.234) and the overall mean person fit residual was 0.008 (SD 0.893). Since both values met this expectation, this indicated that the items and persons fit the model. The "item-trait interaction" statistic reported that the chi-squared value was non-significant [chi-squared = 1049.33 (0.003); given a Bonferroni adjustment fit level of 0,0005)], meaning that the hierarchical ordering of items was invariant across the trait. DIF was tested for academic grades of students, and it was found that all the items were DIF-free.

When the unidimensionality of the 103-item item bank was examined by PCA, there was no pattern violating this assumption (t=%4.6; CI %3.4-%5.7). When the assumption of local independence was tested, there was no pair of items that had a residual correlation of 0.30 or more. For the person-item threshold distribution, person and item locations were logarithmically transformed and plotted on the same continuum.

Figure 2 shows person and item locations on the x-axis. Figure 2 also demonstrates that the item bank was well targeted with the mean of the persons at 0.597 on the logit scale, and few people were outside of the operational range of the scale. As seen in the graphic, the person distribution (top of the figure) was well matched by the item distribution (bottom of the figure).

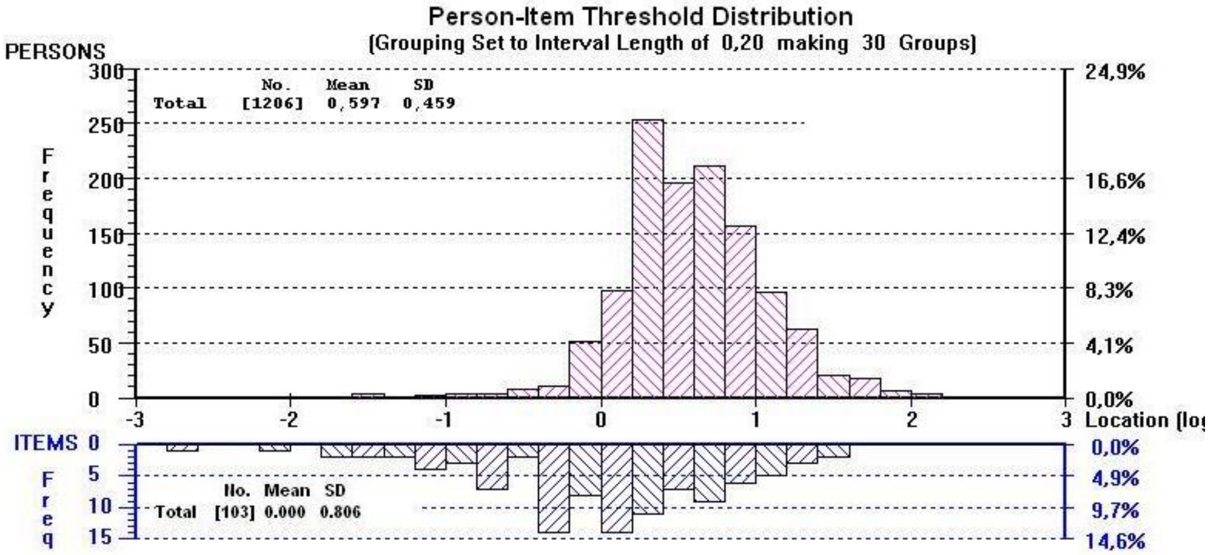


Figure 2. Person-item threshold distribution (103 items)

PSI and KR-20 of the item bank were computed as 0.77. Since the threshold of acceptance for PSI is 0.7, the computed value indicated that it is possible to statistically differentiate between two groups of respondents. This result showed that the items effectively separated the persons.

3.2. Content validity of item bank

When experts from several medical specialties and measurement and evaluation experts examined the content validity of the item bank, they concluded that the item bank contained enough questions for a representative and balanced sampling of the prescribed blueprint.

3.3. Simulation analysis

In this study, simulated CAT application was conducted to evaluate the agreement between θ_{CAT} and θ_{RM}.

3.4. Simulated CAT application [N (0,1)]

The number of items used in CAT, which was carried out with responses derived from 1000 individuals from the distribution of mean with 0 and variance with 1, ranges between 11 and 45 for various standard error levels as shown in Table 2.

Table 2. Descriptive Statistics for Number of Items Administered in CAT application and Correlation and Agreement between *θ_{RM} and θ_{CAT}estimations [N (0,1)]*

Stopping rule	Mean number of items (±SD) [Median (min-max)] usedin CAT	r	ICC (95% CI)
Standard Error: 0.30 *Reliability: 0.90*	45 (±3) [44 (43-50)]	0.975**	0.989 [0.988-0.990]
Standard Error:0.40 *Reliability: 0.84*	24 (±3) [23 (23-50)]	0.940**	0.971 [0.967-0.974]
Standard Error: 0.50 *Reliability: 0.75*	14 (±0.8) [14 (13-21)]	0.886**	0.941 [0.933-0.948]
Standard Error:0.548 *Reliability: 0.70*	11 (±0.53) [11 (11-16)]	0.868**	0.928 [0.919-0.937]

SD: Standard Deviation, r: Correlation Coefficient, ICC: Intraclass Correlation Coefficient., CI: Confidence Interval,
**: p<0.001

In CAT applications, an estimation with a reliability of 0.75 can be made using 14 questions (reduced by % 86,4). When compared to paper and pencil tests (based on all items in the bank), CAT resulted in a 56.3-88.5% decrease in the number of items.

The research findings illustrated that there is a high (r=0.868-0.975 and ICC=0.928-0.989, respectively) correlation and agreement (for standard error 0.3, 0.4, 0.5, 0.548) between θ estimations obtained from paper-and-pencil (θ_{RM}) and CAT applications (θ_{CAT}) and these findings are statistically significant. Ninety-five percent ranges of agreement between θ_{CAT} and θ_{RM} according to the Bland-Altman approach were -0.39 to 0.39, -0.64 to 0.65, -0.83 to 0.83, and -0.92 to 0.96 when the stopping rule was set to standard errors of 0.30, 0.40, 0.50, and 0.548, respectively. In addition, 942 of 1000, 953 of 1000, 960 of 1000, and 935 of 1000 converged estimates were within the 94-96% agreement limits for different standard error, respectively (Figure 3).

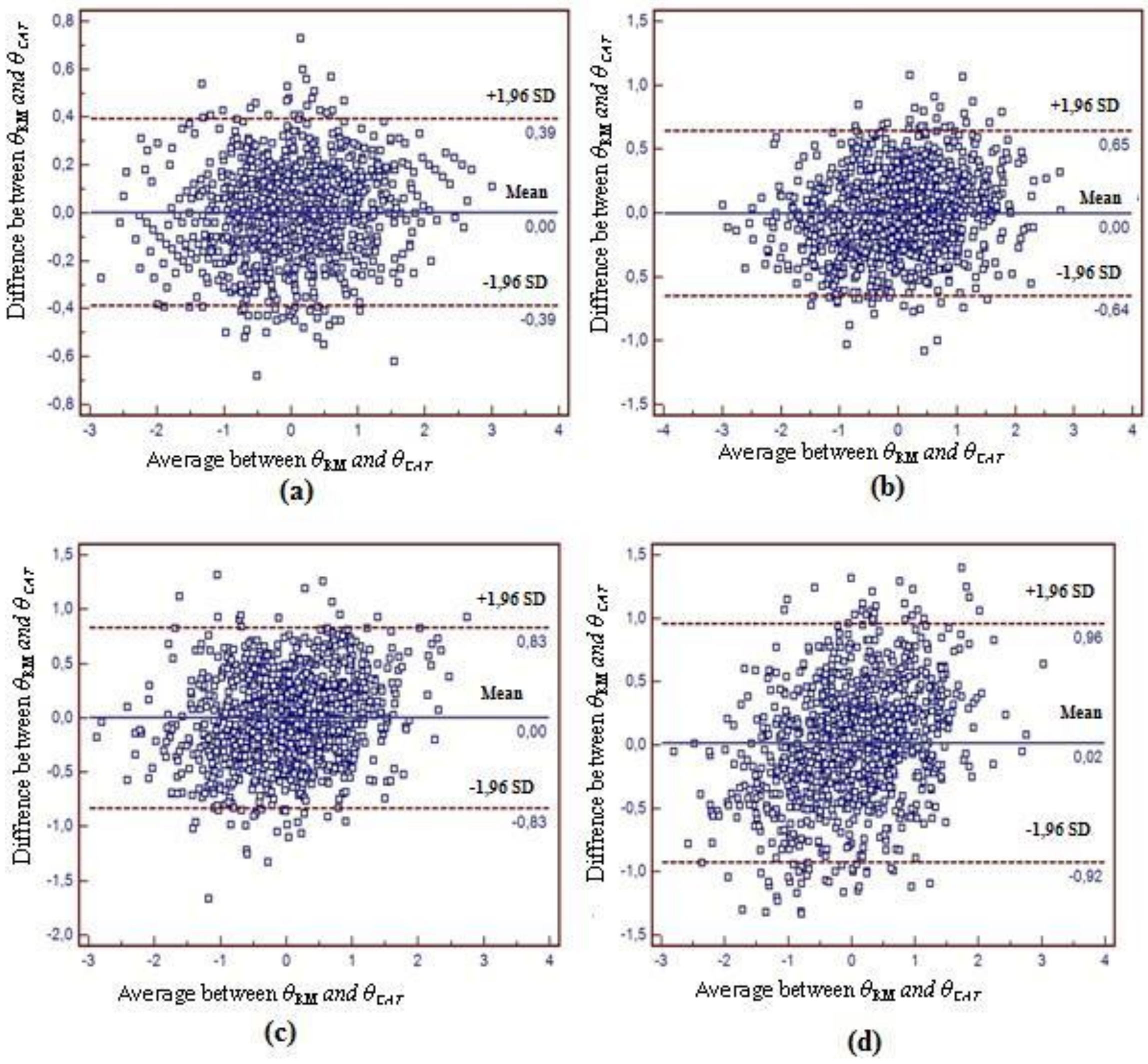


Figure 3. Bland-Altman Plot for [N (0,1)] (a) SE=0.30, (b) SE=0.40, (c) SE=0.50, (d) SE=0.54

3.5. Simulated CAT application [N (0,3)]

The number of items used in CAT, which was carried out with responses derived from 1000 individuals from the distribution of mean with 0 and variance with 3, ranges between 12 and 75 for various standard error levels (Table 3).

Table 3. Descriptive Statistics for Number of Items Administered in CAT application and Correlation and Agreement between θ_{RM} *and* θ_{CAT}*estimations* [N (0,3)]

Stopping rule	Mean number of items (±SD) [median (min-max)] usedin CAT	r	ICC (95% CI)
Standard Error: 0.30 *Reliability: 0.90*	75 (±27) [76 (42-103)]	0.998**	0.999 [0.999-0.999]
Standard Error: 0.40 *Reliability: 0.84*	35 (±15) [27 (22-68)]	0.992**	0.995 [0.995-0.996]
Standard Error: 0.50 *Reliability: 0.75*	17 (±3) [15 (15-23)]	0.984**	0.986 [0.984-0.987]

In CAT applications, an estimation with 0.75 reliability can be made using 17 questions (reduced by %83.4). When compared to paper and pencil tests, CAT amounted to a reduction in the number of items administered by 27.6-88.3%.

The research findings illustrated that there is a high (r=0.984-0.998 and ICC 0.928-0.989, respectively) correlation and agreement (for standard error values 0.3, 0.4, 0.5, 0.548) between θ estimations obtained from paper-and-pencil (θ_{RM}) and CAT applications (θ_{CAT}) and these findings are statistically significant.

Ninety-five percent ranges of agreement between θ_{CAT} and θ_{RM} according to Bland-Altman approach were -0.27 to 0.28, -0.57 to 0.56, -0.96 to 0.94, and -1.21 to 1.28 when the stopping rule was set to standard error of 0.30, 0.40, 0.50, and 0.548, respectively. In addition, 921 of 1000, 948 of 1000, 973 of 1000, and 978 of 1000 converged estimates were also within the 94-96% agreement limits for different standard error, respectively (Figure 4).

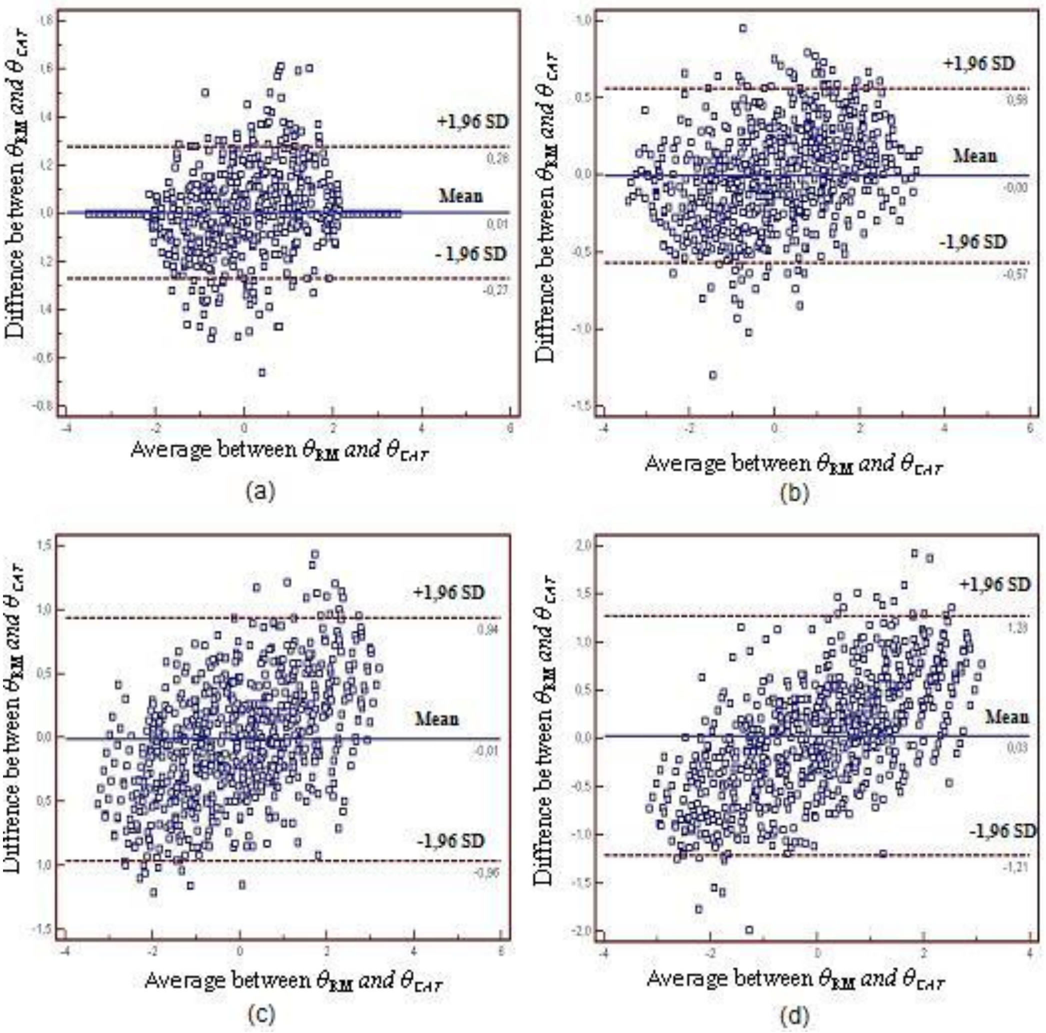


Figure 4. Bland-Altman Plot for [N (0,3)] (a) SE=0.30, (b) SE=0.40, (c) SE=0.50, (d) SE=0.54

4. DISCUSSION and CONCLUSION

This study aimed to demonstrate whether the Rasch model is an alternative to Classical Test Theory, in order to improve the PT in medical education. This study will provide the initiative to investigate the potential for applying CAT in PT.

PT has some disadvantages; firstly, the bulkiness may cause demotivation, boredom, and tiredness. Furthermore, if the questions are too difficult, it could discourage students, while questions that are too easy could be uninteresting for students. In addition, items that are too easy or too difficult do not give enough information about students' ability. Previous research for CAT usage demonstrated that CAT reduces these disadvantages by shortening test length and providing a flexible test format. CAT application offers students a set of questions that are matched to their ability levels, thus providing the examinees with individualized examinations.

The present findings showed that 103 items formed a unidimensional item bank. The ability of the students estimated by CAT was highly correlated with those of the full item set. The average number of items needed to estimate ability was only 13.6% of the full item set of PT/ paper-pencil based PT (for ([N(0,1) and 0.75 reliability], and 14,6 % of the full item set PT (for ([N(0,3) and 0.75 reliability]. θ_{CAT} and θ_{RM} correlated and agreed well for both populations ([N(0,1) and ([N(0,3)]). This study demonstrated that the test length could be shortened without decreasing reliability.

The follow-up of such an approach raises developmental challenges. In this study, the number of items included in the item bank was less than the average for CAT standards. For CAT application, item banks should contain a large number of items considered important to work on the item bank by the CAT designer. Previous studies have been based on item numbers ranging from less than 100 to several thousand items. In this study, however, the number of items was relatively low, with student ability and item difficulty distribution mirroring each other (as shown in Figure 1), and items distributed across the range of the trait being measured. For this reason, although the item bank for this study was relatively small, the results suggested that CAT worked well.

This study intended to discuss the steps required to build an item bank for PT. At the end of this process, a unidimensional item bank that represented "general medical knowledge" was developed for CAT application. However, PT builds on a blueprint that includes specific subdomains (such as the cardiovascular system or the discipline of anatomy) as a part of the overall domain. Extensive feedback and patterns of knowledge growth within specific subdomains could be provided to students and other stakeholders in addition to overall knowledge build-up. To provide detailed feedback and knowledge growth patterns, the items could be divided into unidimensional subsets, and several item banks could be constructed as a possible solution for PT CAT application. In addition, multidimensional CAT procedures based on multidimensional IRT (MIRT) might be another solution for PT CAT. As a result, a reduction of over 80% of the items in CAT format of PT could test its potential to follow candidates' progress practically through educational programs.

CAT application's utility for medical course assessment has been demonstrated in this study. To the authors' knowledge, there have been no other studies about the CAT application in PT for medical school. It should be emphasized that the purpose of this study was not to investigate whether PT CAT based on 103 items should replace the current paper-pencil based PT. This study aimed to discuss the steps to build an item bank and illustrated the utility of CAT implementation as an example.

This study showed that it is possible to develop an appropriate item bank for the PT using the Rasch model, and that the difficulty administering large number of items in PT can be reduced by CAT application. The results of this study will encourage the implementation of CAT in medicine and in other disciplines.

Acknowledgements

This article is extracted from doctorate dissertation entitled "Developing an Item Bank for Progress Test and Application of A Computerızed Adaptive Testing by Simulation in Medical Education" (Advisor: Prof. Dr Nizamettin Koç, Co-advisor Atilla Halil Elhan PhD Dissertation, Ankara University Institute of Educational Sciences, Ankara/Turkey, 2013)

This study has previously been presented in *International 8th Statistics Congress (ISC2013), Antalya, 2013.*

Conflicts of Interest

The authors report no conflicts of interest. The authors alone areresponsible for the content and writing of this article.

ORCID

Aysen Melek AYTUG KOSAN https://orcid.org/0000-0001-5298-2032
Nizamettin KOÇ https://orcid.org/0000-0002-3308-7849
Atilla Halil ELHAN https://orcid.org/0000-0003-3324-248X
Derya ÖZTUNA https://orcid.org/0000-0001-6266-3035

5. REFERENCES

Abberger, B., Haschke, A., Wirtz, M., Kroehne, U., Bengel, J., & Baumeister, H. (2013). Development and evaluation of a computer adaptive test to assess anxiety in cardiovascular rehabilitation patients. *Archives of Physical Medicine and Rehabilitation, 94*(12), 2433-2439. Doi: 10.1016/j.apmr.2013.07.009

Andrich, D. (1988). Rasch models for measurement. The USA: Sage Publications Inc.

Andrich D, Lyne A, Sheridan B, Luo G. RUMM2020. Perth: RUMM Laboratory Pty Ltd. 2003 Freeman, A., Van Der Vleuten, C., Nouns, Z., & Ricketts, C. (2010). Progress testing internationally. *Medical Teacher, 32*(6), 451-455. Doi: 10.3109/0142159X.2010.485231

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics, 37*(3), 387-416. Doi: 10.3102/1076998611411913

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *BMJ, 310*(6973), 170. Doi: 10.1136/bmj.310.6973.170

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307-310. Doi: 10.1016/S0140-6736(86)90837-8

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135-160. Doi: 10.1177/096228029900800204

Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research, 16*(1), 95-108. Doi: 10.1007/s11136-007-9168-6

Elhan A. H., Küçükdeveci A. A., & Tennant A. (2010). The rasch measurement model. Franchignoni F. (Ed.) *Research issues in Physical & Rehabilitation Medicine*. Advances in Rehabilitation. Maugeri Foundation 19, 89-102

Fisher, W. P. (1992). Reliability statistics. *Rasch Measurement Transactions, 6*(3), 238.

Freeman, A., Van Der Vleuten, C., Nouns, Z., & Ricketts, C. (2010) Progress testing internationally. *Medical Teacher*. 2010. *32*(6), 451-455. Doi: 10.3109/0142159X.2010.485231

Hambleton, R. K. (1991). *Fundamentals of item response theory*. The USA: Sage publications.

Linacre, J. M. (2000). Computer adaptive testing: A methodology whose time has come. Chae, S.-Kang, U. Jeon, E. Linacre, JM (eds.): *Development of Computerised Middle School Achievement Tests,* MESA Research Memorandum.

Marais, I., & Andrich, D. (2007). *RUMMss. Rasch unidimensional measurement models simulation studies software*. The University of Western Australia, Perth.

Nunnally, J., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Öztuna, D. (2008). Kas iskelet sistem sorunlarının özürlülük değerlendiriminde bilgisayar uyarlamalı test yönteminin uygulanması (Implementing computer adaptive testing method to estimate disability levels in musculoskeletal system disorders). (Doctoral Dissertation). Ankara Üniversitesi Sağlık Bilimleri Enstitüsü. Ankara

Öztuna D. (2012). *A computerized adaptive testing software (CAT): SmartCAT*. European Rasch Training Group (ERTG) Meeting, 17-19 April 2012, Leeds, UK.

Pallant, J. F., & Tennant, A. (2007). An introduction to the rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, *46*(1), 1-18. Doi: 10.1348/014466506X96931

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420. Doi: 10.1037/0033-2909.86.2.420

Tavakol, M., & Dennick, R. (2012). Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. *Medical Teacher*, *34*(3), e161-e175. Doi: 10.3109/0142159X.2012.651178

Tennant, A., & Pallant, J. (2006). Unidimensionality matters! (A Tale of Two Smiths?). *Rasch Measurement Transactions*, *20*(1), 1048-1051.

Tennant, A., & Conaghan, P. G. (2007). The rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, *57*(8), 1358-1362. Doi: 10.1002/art.23108

Teresi, J. A., Kleinman, M., & Ocepek ‐ Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. *Statistics in Medicine*, *19*(11 ‐ 12), 1651-1683. Doi: 10.1002/(SICI)1097-0258(20000615/30)19:11/12<1651: AID-SIM453>3.0.CO;2-H

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., & Steinberg, L. (2001). Computerized adaptive testing: A primer. *Qual Life Res*, *10*, 733-734. Doi: 10.1023/A:1016834001219

Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement*, *21*(4), 331-345. Doi: 10.1111/j.1745-3984.1984.tb01038.x

Wrigley, W., Van Der Vleuten, C. P., Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. Medical Teacher, 34(9), 683-697. Doi: 10.3109/0142159X.2012.704437

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 670–692
https://dx.doi.org/10.21449/ijate.661803

Published at http://www.ijate.net http://dergipark.org.tr **_Research Article_**

Jamovi: An Easy to Use Statistical Software for the Social Scientists

Murat Doğan Şahin[1] *, Eren Can Aybek[2]

[1] Department of Educational Measurement and Evaluation, Anadolu University, Eskişehir, Turkey
[2] Department of Educational Measurement and Evaluation, Pamukkale University, Denizli, Turkey

ARTICLE HISTORY

Received: 03 August 2019
Revised: 31 October 2019
Accepted: 16 December 2019

KEYWORDS

Jamovi,
R,
Statistical software,
Free software

Abstract: This report aims to introduce the fundamental features of the free Jamovi software to academics in the field of educational measurement for use at undergraduate and graduate level research. As such, after introducing the R based interface and the integrated development environment, the core functions of Jamovi are presented, the installation for GNU7Linux, Windows, and MacOS is explained and screenshots of frequently conducted statistical analyses are provided. Additionally, the module support of Jamovi is presented, along with a use case scenario on developing further functionality for Jamovi using modules. Specifically, conducting meta-analysis and Bayesian statistics using modules in Jamovi are explained through examples.

1. INTRODUCTION

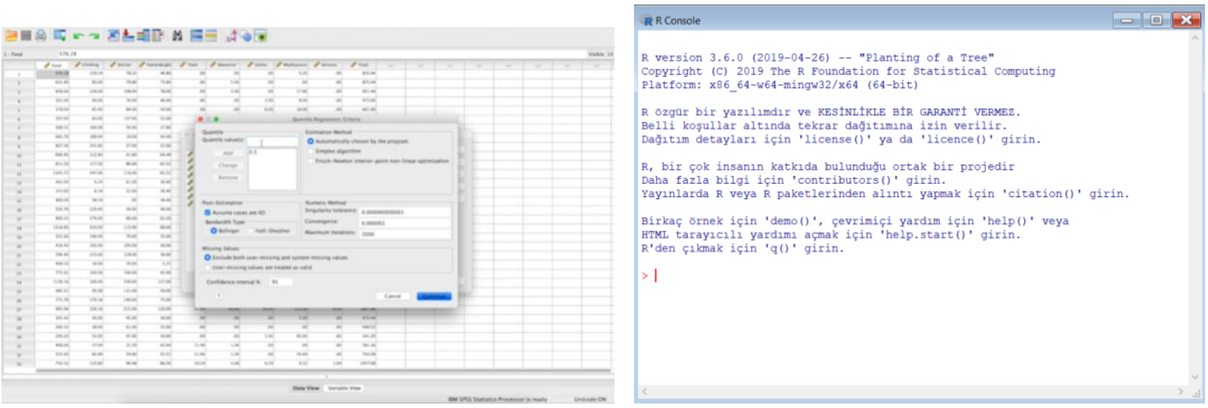
One of the most important stages of scientific research, data analysis, was one conducted by hand and then by calculators. Today, various computer software is used to assist this process. The public access provided to data from wide-scale tests such as the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMMS), along with the rapid increase in the amount of data collected in social sciences have made statistical software an integral part of the quantitative research process.

Muenchen (2019) determined the market shares of statistical software, concluding that IBM’s SPSS Statistics program was the most preferred software in academic studies as of the end of 2018. Meanwhile, R has been referenced half as much as IBM’s SPSS Statistics program in academic articles. The situation may be attributed to the ease of use of SPSS through its graphical user interface, it being used extensively by academics for many years, and SPSS being the primary software used in statistics training and education. Despite lacking the ease of use of SPSS, as the second most popular software in the field, R has gained new functionality through packages published by researchers around the world due to its open source nature. R is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (Formerly AT&T, now Lucent Technologies) by John Chambers and colleagues

CONTACT: Murat Doğan Şahin ✉ mdsahin@anadolu.edu.tr Anadolu University, Faculty of Education, Eskişehir, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

(R Core Team, 2019a). GNU is an acronym for GNU is Not Unix, and is a free operating system developed by Richard Stallman in 1983 (GNU, 2019). While R is free software developed based on this operating system, today it is capable of running on common operating systems such as GNU/Linux, Microsoft Windows, and Mac OS X. Developed on the free and open software philosophy of GNU, anyone may download and use R, study its code, modify it and distribute their modifications (R Core Team, 2019a). Screenshots of IBM SPSS Statistics v.20 and R may be compared in Figure 1.



```
R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R özgür bir yazılımdır ve KESİNLİKLE BİR GARANTİ VERMEZ.
Belli koşullar altında tekrar dağıtımına izin verilir.
Dağıtım detayları için 'license()' ya da 'licence()' girin.

R, bir çok insanın katkıda bulunduğu ortak bir projedir
Daha fazla bilgi için 'contributors()' girin.
Yayınlarda R veya R paketlerinden alıntı yapmak için 'citation()' girin.

Birkaç örnek için 'demo()', çevrimiçi yardım için 'help()' veya
HTML tarayıcılı yardımı açmak için 'help.start()' girin.
R'den çıkmak için 'q()' girin.

> |
```

(a) IBM SPSS Statistics v26 for OS X

(b) R v.3.6.0 (Console) for Windows

Figure 1. Screenshots for IBM SPSS Statistics and R
(a) Source URL: https://www.ibm.com/downloads/cas/RJBNRBVB

As can be seen in Figure 1, while R only provides the user with a console screen, IBM statistics allows users to conduct any analysis they want (albeit any analysis the developers allow) with a few mouse clicks. While R may only have a console, it can download a package developed by a researcher from another part of the world with a simple *install.packages()* function and conduct any analysis. The analyses users may conduct in R are 'limited' to the more than 15000 packages developed for R (R Core Team, 2019b). Additionally, users can develop their own packages and offer them for use by other researchers around the world. In this regard, while SPSS is a user-friendly program with limited functionality, R appears limitless but cumbersome in comparison. Considering researchers, especially those in social sciences, may not be skilled at coding, it's easy to understand the reasoning behind the proliferation of IBM SPSS Statistics.

1.1. Statistical Software and Resources

The prevalent use of SPSS by academics allows them to transfer this use to future generations, namely students. 11 of the results on the first page of an Amazon.com search for books on *statistics for social science* are for IBM SPSS Statistics, while only two are results pertaining to R (Amazon, 2019). This situation is indicative of the prevalence of SPSS in statistics education. Teaching the use of proprietary software to undergraduate and graduate students encourages future professionals and researchers to use this proprietary software.

1.2. R and Its Graphical User Interfaces / Integrated Development Environments

As can be seen in Figure 1, the interface of R fundamentally consists merely of a screen terminal. While functions such as package loading, package retrieval, and statistical analysis is possible through the terminal, it is clearly not user-friendly. Therefore, various interfaces and integrated development environments aiming to ease the use of R were developed. RKWard, RStudio, JASP, and Jamovi are a few of them.

RKWard is a comprehensive interface developed for R (RKWard, 2019). It provides access to the R terminal and allows for point-and-click analyses through its interface. The Spreadsheet function allows variable definition and data entry, correlation, regression and average comparison along with item response theory analyses. RKWard is free software and a screenshot is presented in Figure 2.

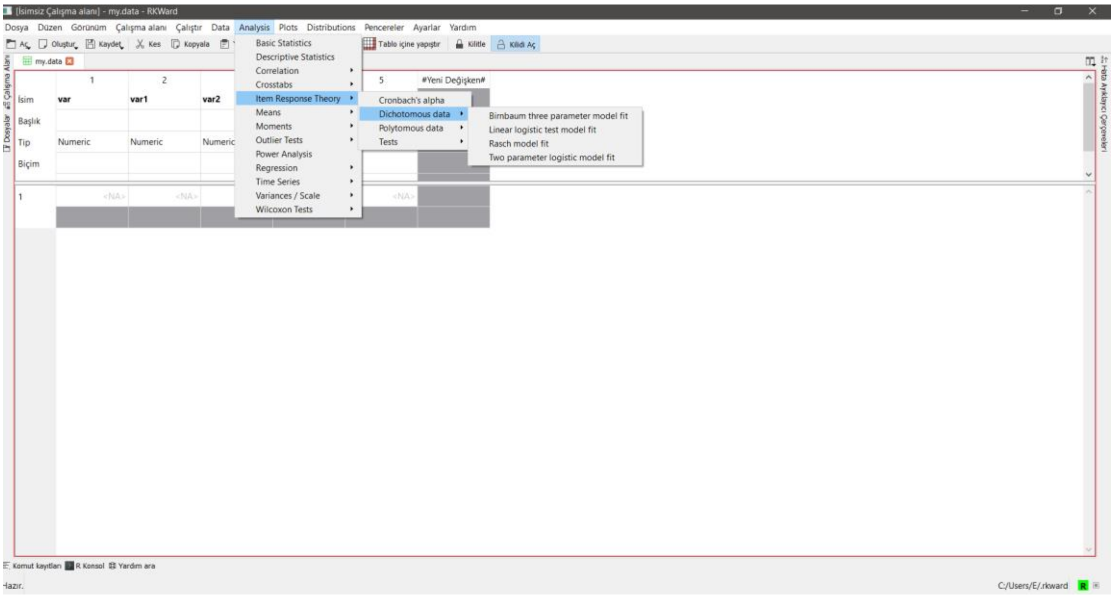
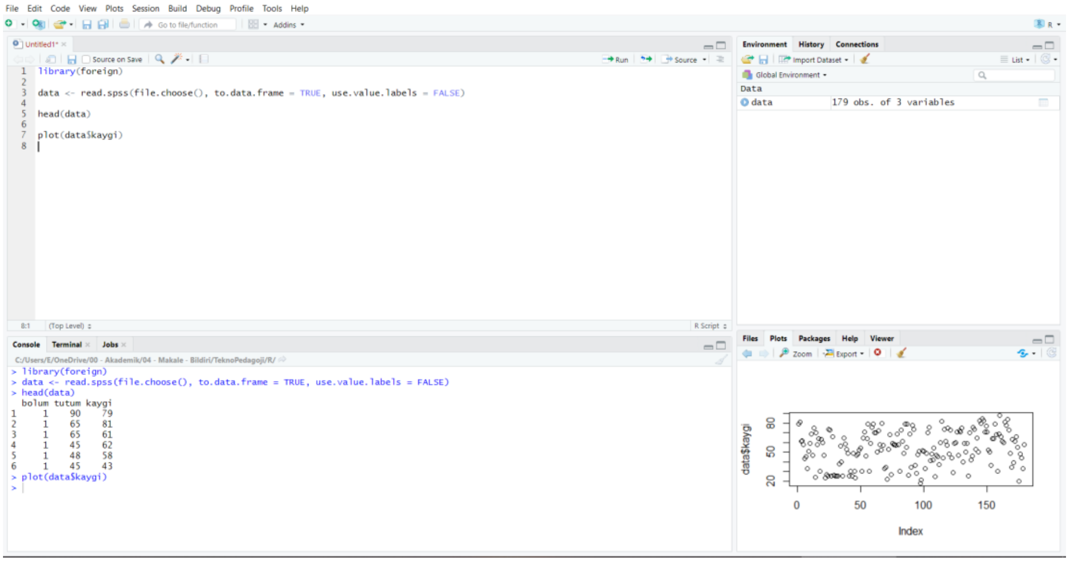

[İsimsiz Çalışma alanı] - my.data - RKWard
Dosya Düzen Görünüm Çalışma alanı Çalıştır Data Analysis Plots Distributions Pencereler Ayarlar Yardım
Aç Oluştur Kaydet Kes Kopyala Tablo içine yapıştır Kilitle Kilidi Aç
my.data
İsim var var1 var2
Başlık
Tip Numeric Numeric Numeric
Biçim
#Yeni Değişken#
<NA>
Basic Statistics
Descriptive Statistics
Correlation
Crosstabs
Item Response Theory
Means
Moments
Outlier Tests
Power Analysis
Regression
Time Series
Variances / Scale
Wilcoxon Tests
Cronbach's alpha
Dichotomous data
Polytomous data
Tests
Birnbaum three parameter model fit
Linear logistic test model fit
Rasch model fit
Two parameter logistic model fit
Komut kayıtları R Konsol Yardım ara
Hazır.
C:/Users/E/.rkward

Figure 2. Screenshot for RKWard 0.7.0b

RStudio is an Integrated Development Environment (IDE) that has become synonymous with R (RStudio, 2019). The greatest advantage of RStudio is the easy access it provides to the objects being studied and the graphs created in R. Additionally, it simplifies working with R scripts. Furthermore, RStudio has simplified creating project folders and version tracking through git while providing the possibility of writing cleaner R code through plugins. The basic packages developed for data science in R are gathered under tidyverse in RStudio. The Shiny package, which allows the creation of a web interface for R code, is also developed by RStudio. A screenshot for RStudio is presented in Figure 3.



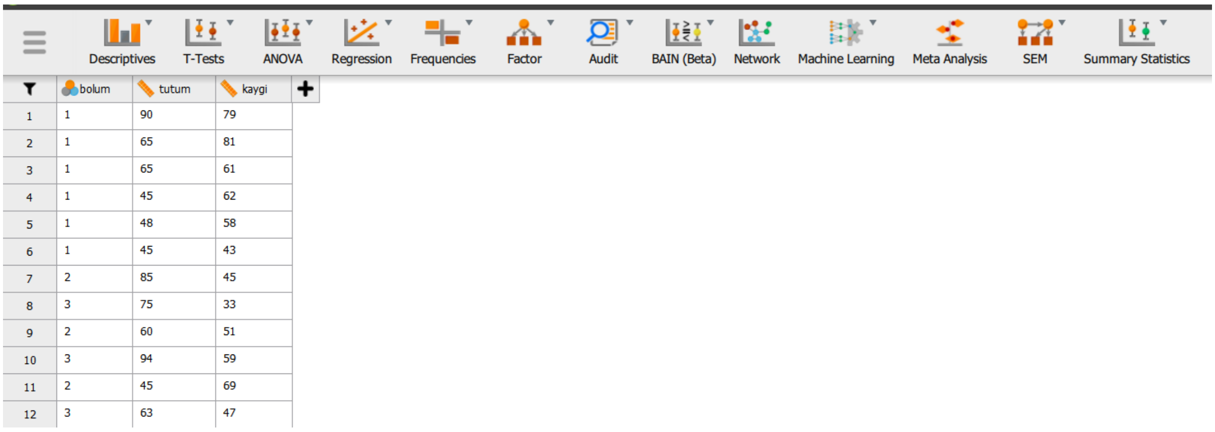
```
library(foreign)

data <- read.spss(file.choose(), to.data.frame = TRUE, use.value.labels = FALSE)

head(data)

plot(data$kaygi)
```

```
data        179 obs. of 3 variables
```

```
C:/Users/E/OneDrive/00 - Akademik/04 - Makale - Bildiri/TeknoPedagoji/R/
> library(foreign)
> data <- read.spss(file.choose(), to.data.frame = TRUE, use.value.labels = FALSE)
> head(data)
  bolum tutum kaygi
1     1    90    79
2     1    65    81
3     1    65    61
4     1    45    62
5     1    48    58
6     1    45    43
> plot(data$kaygi)
>
```

Figure 3. Screenshot for RStudio

JASP is a simple interface software developed on R (JASP Team, 2019). JASP stands out with its support for Bayesian methods, allowing the user to conduct Bayesian t-tests and Bayesian ANOVA with a few clicks. Additionally, JASP is capable of conducting tests using a frequentist approach and has gained powerful functions such as network analysis, machine learning algorithms, meta-analysis, and structural equation modeling through current modules. JASP supports many data formats such as cxv, xls, and sav, and allows for data analysis but does not allow for data manipulation. JASP runs on GNU/Linux, Windows, and MacOS, and can also be used online through Rollapp. A screenshot of JASP is provided in Figure 4.



	bolum	tutum	kaygi
1	1	90	79
2	1	65	81
3	1	65	61
4	1	45	62
5	1	48	58
6	1	45	43
7	2	85	45
8	3	75	33
9	2	60	51
10	3	94	59
11	2	45	69
12	3	63	47

Figure 4. Screenshot for JASP

Another free software based on R, and the locus of this article, is Jamovi. Detailed information on Jamovi is provided in the sections below.

2. JAMOVI

2.1. Developers

Jamovi is being developed by a team who branched off from JASP developers and started running their own project (The Jamovi Project, 2019a). Jonathan Love, Damian Dropmann and Ravi Selker are members of this team. In their own words; "*we found that our goals and ambitions consistently went beyond their scope, and decided the best way to move forward was to found a new project: the jamovi project*" (The jamovi project, 2019b).

Possibly the most significant point emphasized by the developers is the community-driven aspect. This means that users in the community may add new features to Jamovi by adding modules. Thus, Jamovi does not remain limited to the initiatives taken by the developer team. In this regard, Jamovi adopts a policy highly suited to the free software philosophy.

2.2. Features

Jamovi provides core functions such as data entry and manipulation, rule based data filtering and variable transformation, and computing with variables. Compatible with popular data file formats such as csv, RData, dta, and sav, Jamovi is capable of conducting many single and multiple variable analyses. Among these analyses are descriptive statistics, t-tests, ANOVA, ANCOVA, MANCOVA, linear and logistic regression, exploratory and confirmatory factor analysis, and nonparametric tests.

Additionally, the module support of Jamovi allows developers to add new functionality if the so please. For example, Walrus for robust statistics and MAJOR for meta-analysis are modules developed and added to the Jamovi library. In this regard, Jamovi presents itself as a sufficient, useful, and free tool for academic research in social sciences in addition to undergraduate and graduate level statistics courses.

2.3. Installation

Once the *download* link on the Jamovi web site is clicked, links for two different versions of Jamovi are presented. One of them is *solid*, while the other is the *current* release. While the solid version is more stable, the current version includes the latest features of Jamovi. While both versions are available for Windows and macOS, only the current version is provided for GNU/Linux and ChromeOS. All Jamovi releases are designed for 64-bit operating systems. The installation steps for Jamovi on GNU/Linux, Windows, and macOS are provided below. As the steps taken are very similar to GNU/Linux, the instructions for ChromeOS are not included in this report.

2.3.1. *Installing on GNU/Linux*

Jamovi for GNU/Linux distributions may be installed through Flatpak. Flatpak is a tool for compiling and distributing desktop applications in GNU/Linux distributions. Multiple programs may be installed with a single command through Flatpak. For installation, Flatpak must be installed on GNU/Linux. Flatpak may be installed by following the instructions at https://flatpak.org/setup/. There are instructions for multiple GNU/Linux distributions at this address. Within this article, installation was conducted on Pop_OS! 19.10, a GNU/Linux distribution based on Ubuntu.

The following terminal command is entered for the installation of Flatpak:

```
sudo apt install flatpak
```

Following this command, the superuser (administrator) password will be requested for the sudo operation. Once the password is entered, installation begins. The Flatpak installation is confirmed through the dialog by pressing Y and the installation of Flatpak is completed. To install Flatpak programs, the Flatpak repositories need to be added. This is achieved with the following terminal command:

```
flatpak remote-add --if-not-exists flathub
https://flathub.org/repo/flathub.flatpakrepo
```

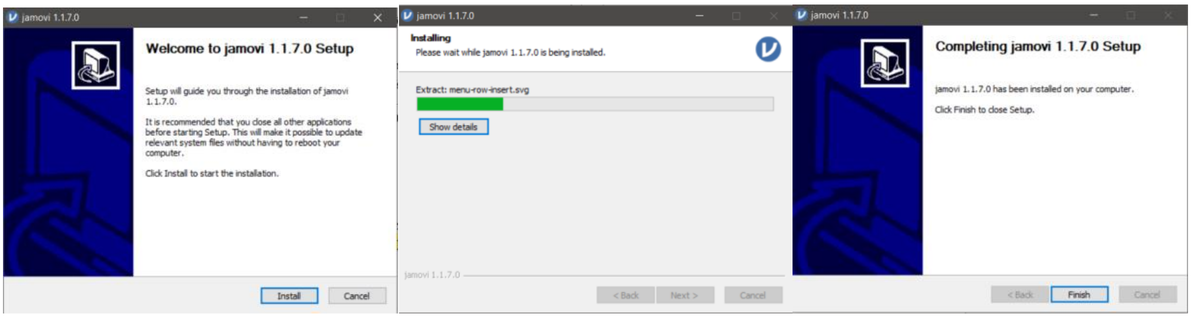
Once the computer is restarted after this, the Terminal is opened once again and the following command is entered:

```
flatpak install flathub org.jamovi.jamovi
```

The dialogs are confirmed by pressing Y, and once the necessary files for Jamovi are downloaded, it is installed on the computer.

2.3.2. *Installing on Windows*

Once the desired version (*solid* or *current*) is downloaded, double clicking on the installation file initiates the installation. Once the *Install* button is clicked, Jamovi is installed and completed by clicking the *Finish* button. The three screens that appear during installation are presented in Figure 5.



Welcome to jamovi 1.1.7.0 Setup

Setup will guide you through the installation of jamovi 1.1.7.0.

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Install to start the installation.

Installing

Please wait while jamovi 1.1.7.0 is being installed.

Extract: menu-row-insert.svg

Completing jamovi 1.1.7.0 Setup

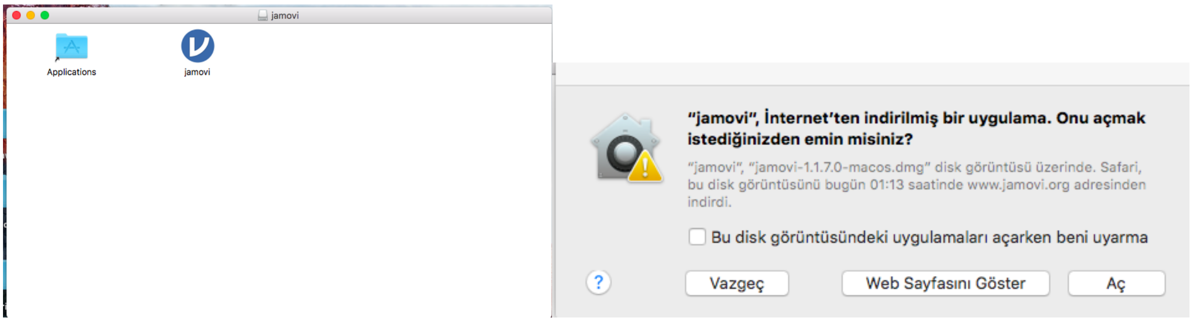
jamovi 1.1.7.0 has been installed on your computer.

Click Finish to close Setup.

Figure 5. Jamovi Windows Installation Screens

2.3.3. *Installing on macOS*

Similar to the Windows installation, the desired version is downloaded. Double clicking on the Jamovi image within the downloaded file opens the program in a new window. This process is presented in Figure 6.



"jamovi", İnternet'ten indirilmiş bir uygulama. Onu açmak istediğinizden emin misiniz?

"jamovi", "jamovi-1.1.7.0-macos.dmg" disk görüntüsü üzerinde. Safari, bu disk görüntüsünü bugün 01:13 saatinde www.jamovi.org adresinden indirdi.

□ Bu disk görüntüsündeki uygulamaları açarken beni uyarma

Figure 5. Jamovi macOS Installation Screens

2.4. First Look

Following installation, Jamovi can be run by double clicking the shortcut icon. Additionally, it can be run from the terminal in GNU/Linux with the `flatpak run org.jamovi.jamovi` command. The first screen to be encountered in Jamovi is presented in Figure 7.

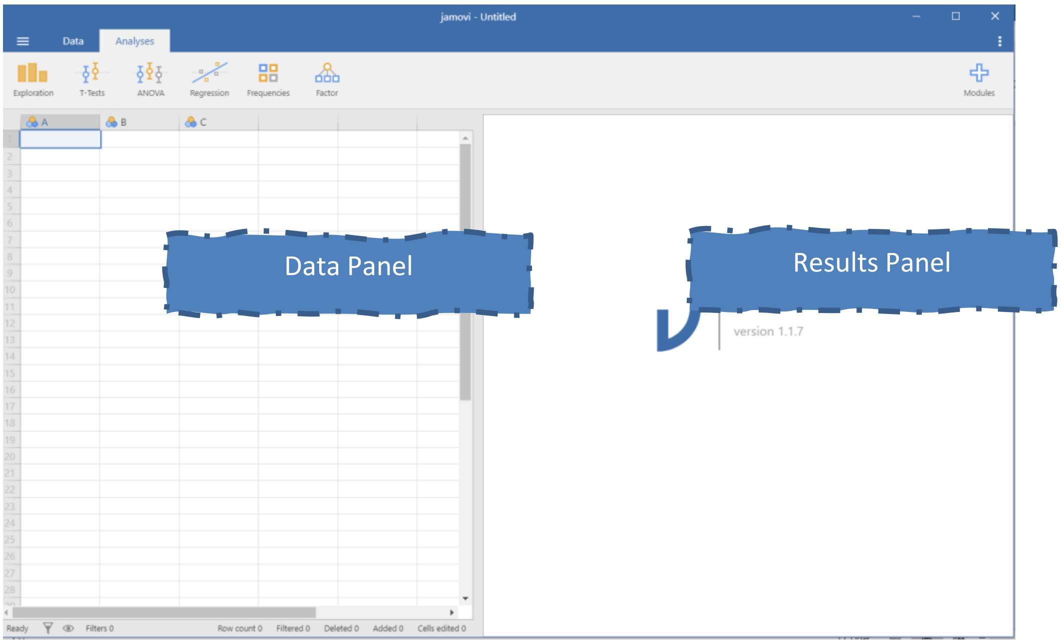


Figure 7. Jamovi Initial Screenshot

Jamovi has *Data* and *Analyses* tabs at the top left corner. The *Data* tab provides processes regarding variables, while the *Analyses* tab is used to execute data analysis processes. The window is divided basically into two panels. The data is presented in the left panel, while the analysis results are presented in the right. When data analysis is to be conducted, a menu regarding the analysis is presented instead of the data panel. The options in the *data* tab are presented in Figure 8.

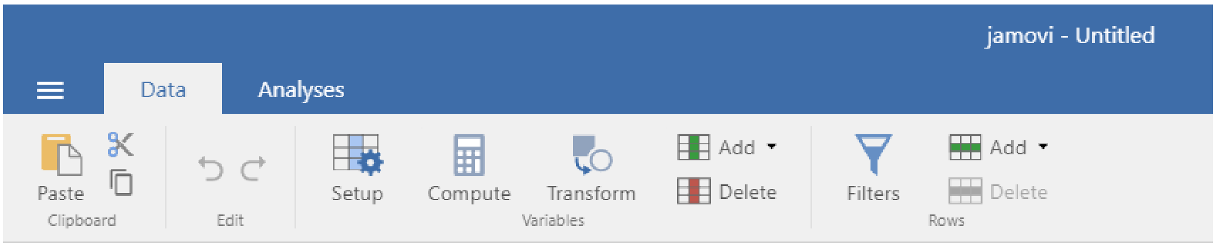


Figure 8. Jamovi Data Tab

Within the data tab, data pasting/copying, variable definition, calculation and transformation, variable addition and removal, and rule-based filtering can be conducted. A detailed image of the menus in the *Analyses* tab are presented in Figure 9.

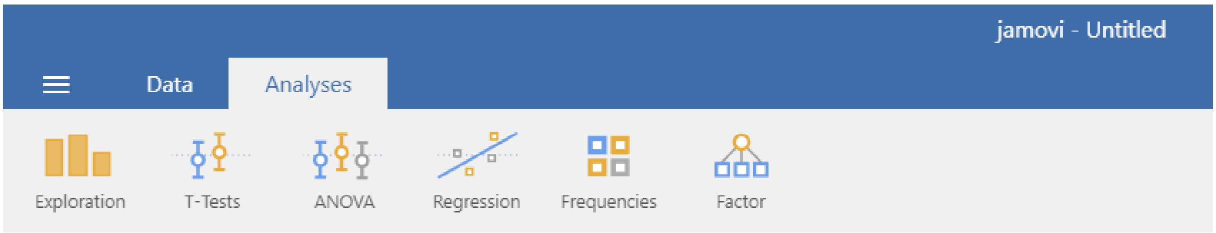
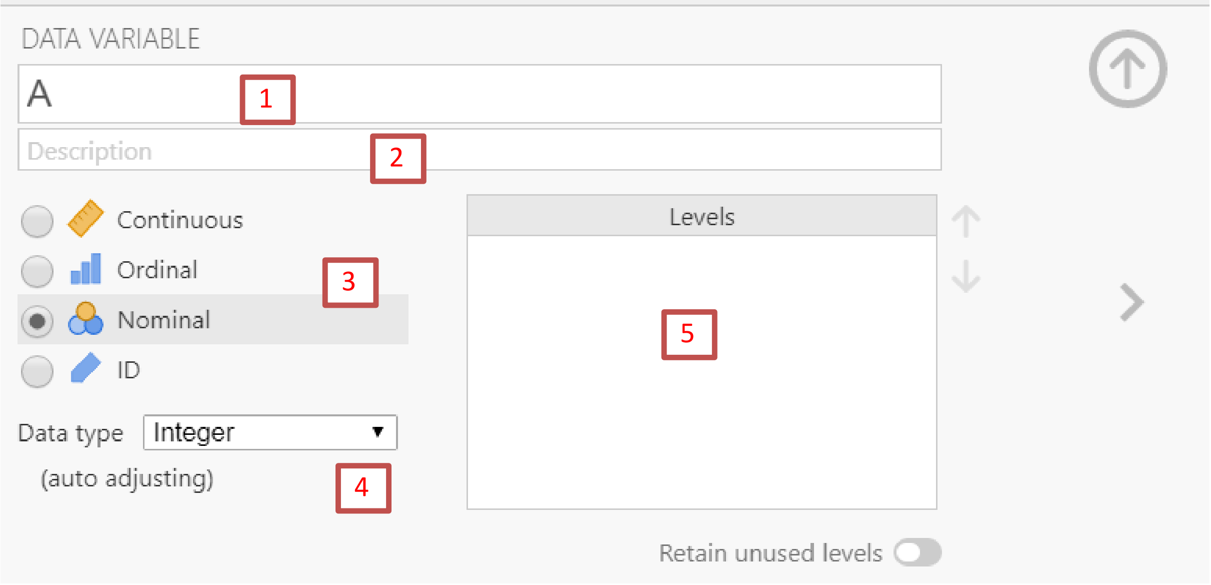


Figure 9. Jamovi Analyses Tab

In this tab, the default analyses available upon the initial installation of Jamovi are listed. New analyses added through modules are added to this section. By default, Jamovi is capable of conducting *descriptive statistics, independent samples t-test, paired samples t-test, one sample t-test, Mann-Whitney U test, Wilcoxon signed rank test, one way ANOVA, repeated measures ANOVA, ANCOVA, MANCOVA, Kruskal-Wallis H test, Friendman test, Correlation, Linear regression, Logistic regression, χ^2 Goodness of fit, χ^2 test for association, McNemar test, Log-Linear regression, reliability analysis, Principal Component Analysis, Exploratory Factor Analysis* and *Confirmatory Factor Analysis*. Additionally, the assumptions of these tests may be tested using Jamovi.

2.5. Data Setup

Prior to data entry, the data must be defined. To do this, the *Setup* button under the *Data* tab is clicked and the screen presented in Figure 10 appears.



DATA VARIABLE

A

Description

○ Continuous
○ Ordinal
● Nominal
○ ID

Data type Integer ▾
(auto adjusting)

Retain unused levels

Figure 10. Data Setup Menu

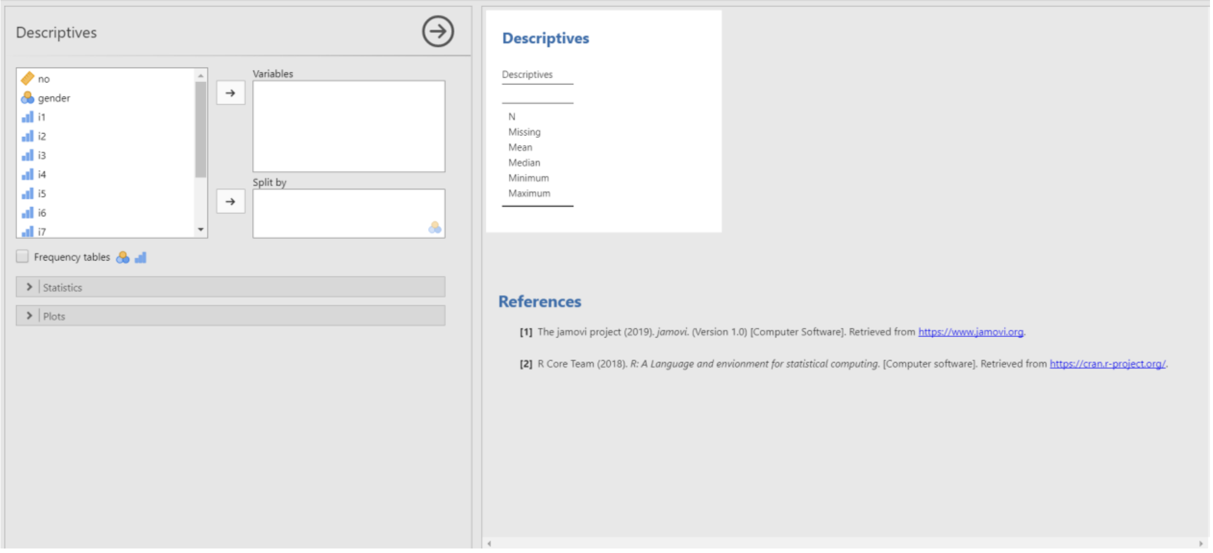
In this screen, the name of the variable is defined in the section marked *1*, while an explanation regarding this variable may be entered in the field marked *2* if desired. The level of measurement for the variable may be defined in the field marked *3*. What is interesting is that in addition to the continuous, ordinal, and nominal options, an option labelled ID is presented. This feature is useful for when participants of a study are assigned ID numbers and greatly eases data entry. The nature of the data to be entered, whether it be *integer*, *decimal*, or *text*, is specified through field *4*. The field marked *5* is used to define the levels of *nominal* variables.

2.6. Statistical Analyses

The statistical analyses that Jamovi is capable of executing by default were presented in section *2.4 First Look*. The focus of this section is on how statistical analyses are conducted in Jamovi in detail, along with examples for descriptive statistics, independent samples t-tests, dependent samples t-tests, one-way ANOVA, correlation, linear regression, and exploratory and confirmatory factor analysis.

2.6.1. *Descriptive Statistics*

To conduct descriptive statistics in Jamovi, the *Descriptives* option is selected from the *Exploration* menu. The screen presented in Figure 11 appears following these actions.



Descriptives

Descriptives

N
Missing
Mean
Median
Minimum
Maximum

References

[1] The jamovi project (2019). *jamovi*. (Version 1.0) [Computer Software]. Retrieved from https://www.jamovi.org.

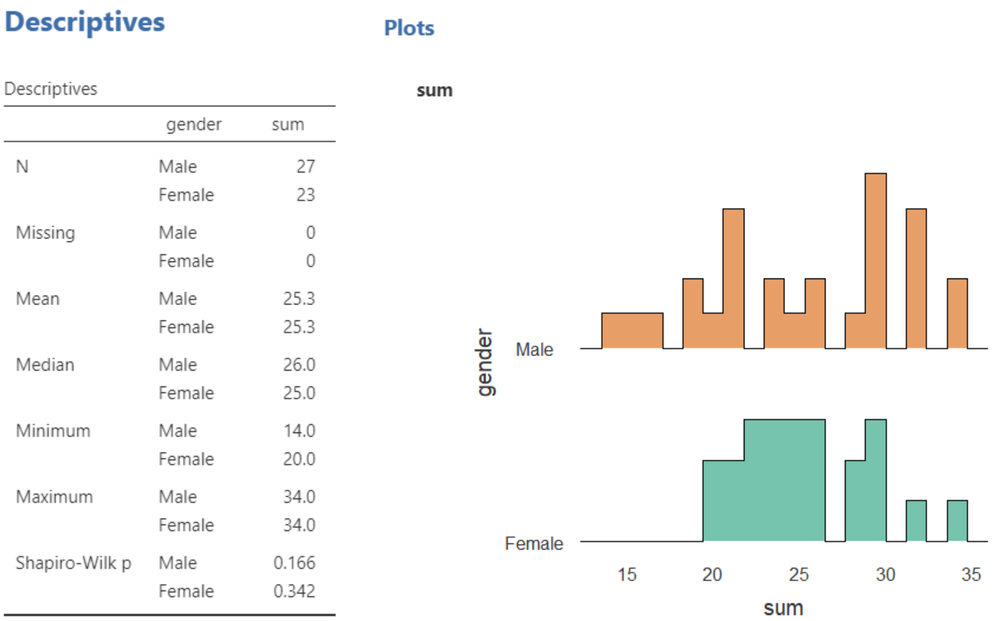
[2] R Core Team (2018). *R: A Language and envionment for statistical computing.* [Computer software]. Retrieved from https://cran.r-project.org/.

Figure 11. Descriptive Statistics Menu

As can be seen, the variables and analysis options are on the left, while the analysis results are on the right. Additionally, a reference section is presented for researchers to refer to this program in the studies in the APA style.

On this screen, once the *Statistics* menu is opened; *N*, *missing value*, *Quartiles*, *standard deviation*, *variance*, *range*, *minimum*, *maximum*, *standard error*, *mean*, *median*, *mode*, *sum*, *skewness*, *kurtosis* statistics along with the *Shapiro-Wilk normality test* are presented. Under the *Plots* menu, the options presented are *histogram*, *density*, *Q-Q*, *box-plot*, *violin* and *bar plots*.

As an example, if one were to require the descriptive statistics of total score based on sex, the *sum* variable is moved to *Variables*, while the *gender* variable is moved to the *split by* section. This is then followed by selecting the *histogram* graph and the *Shapiro-Wilk* test. The results of this analysis are presented in figure 12.



Descriptives

Descriptives

	gender	sum
N	Male	27
	Female	23
Missing	Male	0
	Female	0
Mean	Male	25.3
	Female	25.3
Median	Male	26.0
	Female	25.0
Minimum	Male	14.0
	Female	20.0
Maximum	Male	34.0
	Female	34.0
Shapiro-Wilk p	Male	0.166
	Female	0.342

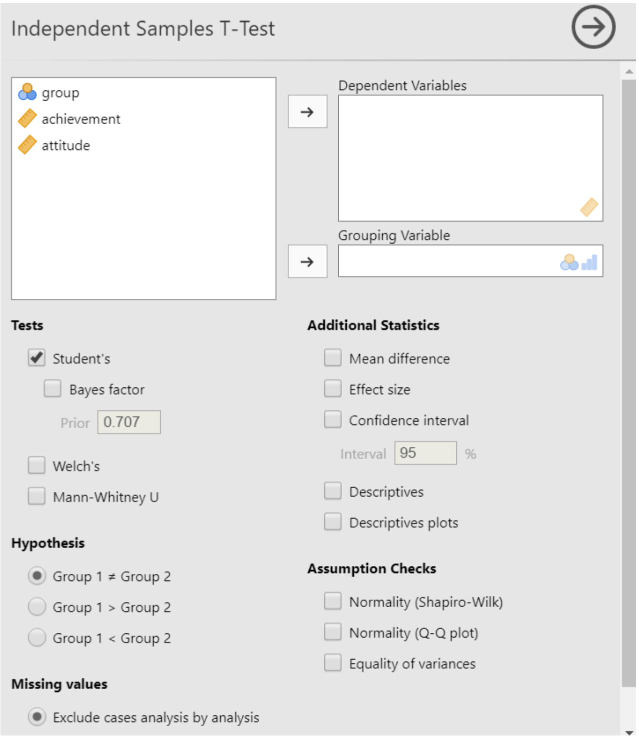
Plots

Figure 12. Descriptive Statistics

As can be seen from the figure, the descriptive statistics are calculated and presented separately based on gender.

2.6.2. *Independent Samples t Test*

Indepentent samples t-tests, which are conducted to compare the arithmetic means of two indepentent samples (Field, 2013) may be executed by clicking on the *Independent Samples T Test* option under the *T-Tests* menu. This menu is presented in Figure 13.



Independent Samples T-Test

group
achievement
attitude

Dependent Variables

Grouping Variable

Tests

☑ Student's

□ Bayes factor

Prior 0.707

□ Welch's

□ Mann-Whitney U

Hypothesis

● Group 1 ≠ Group 2

○ Group 1 > Group 2

○ Group 1 < Group 2

Missing values

● Exclude cases analysis by analysis

Additional Statistics

□ Mean difference

□ Effect size

□ Confidence interval

Interval 95 %

□ Descriptives

□ Descriptives plots

Assumption Checks

□ Normality (Shapiro-Wilk)

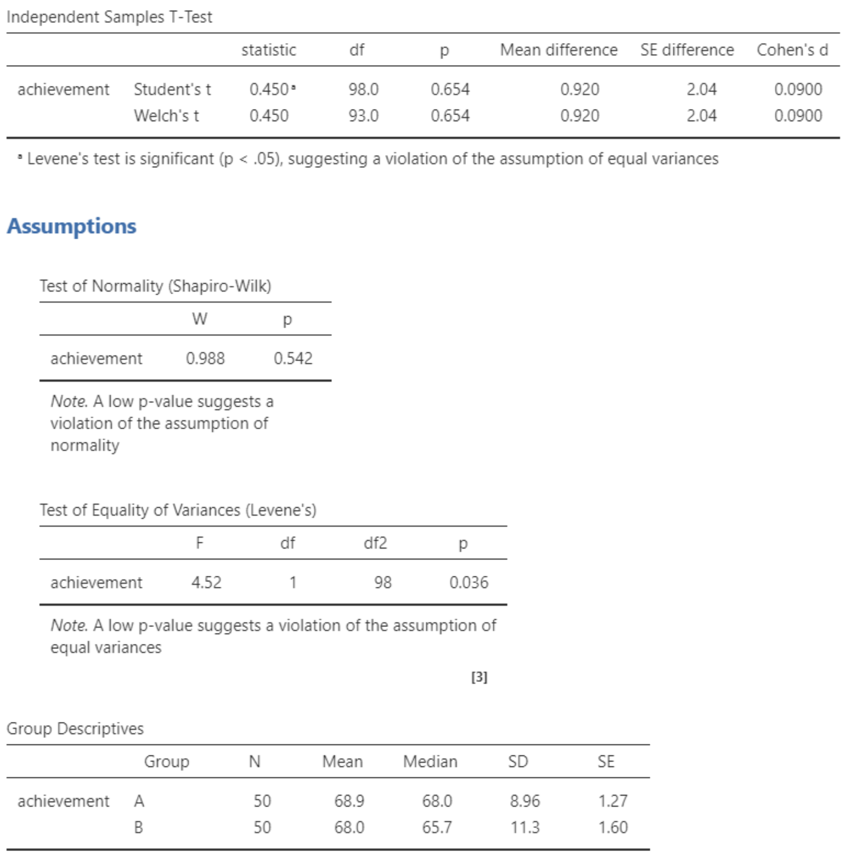
□ Normality (Q-Q plot)

□ Equality of variances

Figure 13. Independent Samples t Test Menu

The dependent variable is assigned to the dependent variables field, while the independent variable is assigned to the grouping variable field. The options for the test include a Student t-test for equal variance, a Welch test for unequal variance, and a Mann-Whitney U test as a nonparametric test. Analysis results may be obtained based on two-way or one-way hypotheses, and tests such as the effect size, descriptive statistics, normality, and variance equality may all be requested under a single menu.

For examples, to determine if there is any significant difference between the academic achievement of students in groups A and B, the findings of the independent samples t-test results are presented in Figure 14.



Independent Samples T-Test

		statistic	df	p	Mean difference	SE difference	Cohen's d
achievement	Student's t	0.450ᵃ	98.0	0.654	0.920	2.04	0.0900
	Welch's t	0.450	93.0	0.654	0.920	2.04	0.0900

ᵃ Levene's test is significant (p < .05), suggesting a violation of the assumption of equal variances

Assumptions

Test of Normality (Shapiro-Wilk)

	W	p
achievement	0.988	0.542

Note. A low p-value suggests a violation of the assumption of normality

Test of Equality of Variances (Levene's)

	F	df	df2	p
achievement	4.52	1	98	0.036

Note. A low p-value suggests a violation of the assumption of equal variances

[3]

Group Descriptives

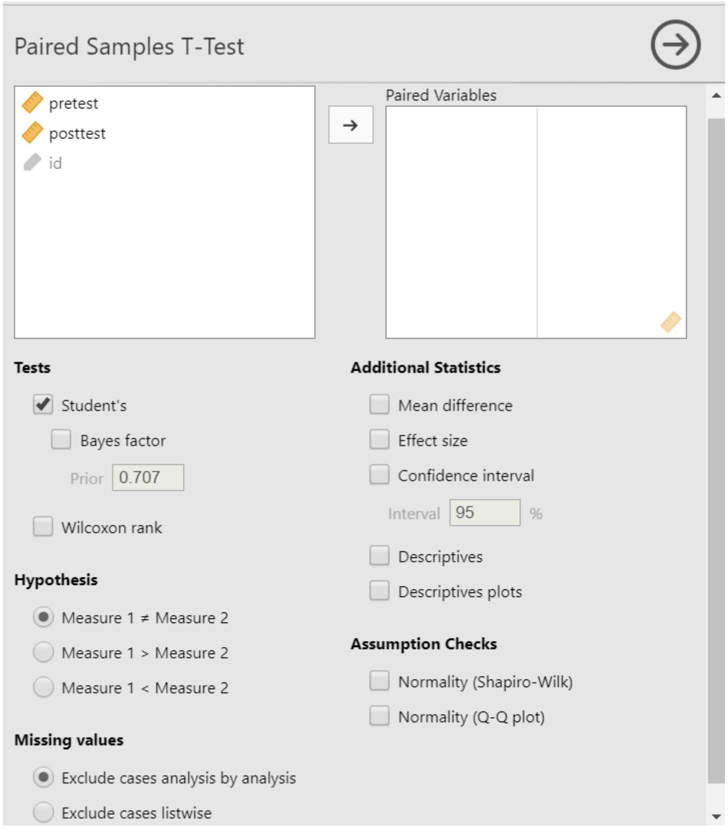
	Group	N	Mean	Median	SD	SE
achievement	A	50	68.9	68.0	8.96	1.27
	B	50	68.0	65.7	11.3	1.60

Figure 14. Independent Samples t-Test Output

As can be seen, the assumption test, test results, and effect size for the independent samples t-test may be requested from the same menu and the outputs may be portrayed in the same section.

2.6.3. *Paired Samples t Test*

The comparison of the means of two dependent measurements are conducted using a paired samples t-test (Navarro & Foxcroft, 2019). To this end, under the *T-Tests* button in Jomavi, the *Paired Samples T-Test* path is selected. Jamovi's menu for paired samples t-tests is shown in Figure 15.



Paired Samples T-Test

pretest
posttest
id

Paired Variables

Tests

☑ Student's

□ Bayes factor

Prior 0.707

□ Wilcoxon rank

Hypothesis

● Measure 1 ≠ Measure 2

○ Measure 1 > Measure 2

○ Measure 1 < Measure 2

Missing values

● Exclude cases analysis by analysis

○ Exclude cases listwise

Additional Statistics

□ Mean difference

□ Effect size

□ Confidence interval

Interval 95 %

□ Descriptives

□ Descriptives plots

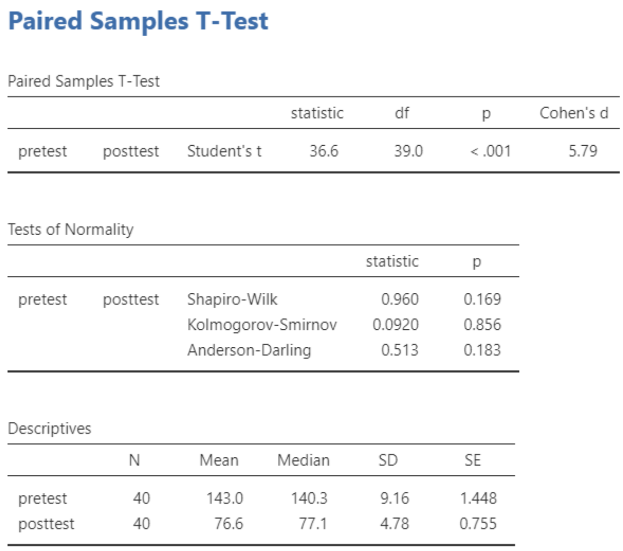
Assumption Checks

□ Normality (Shapiro-Wilk)

□ Normality (Q-Q plot)

Figure 15. Paired Samples t Test Menu

In this menu, the selection between paired samples t-test and the Wilcoxon signed-rank test as a nonparametric test may be done. Additionally, the assumptions to be tested may be determined as either one-way or two-way. Information regarding assumptions such as effect size, descriptive statistics and normality tests may also be requested. The results of a simple paired samples t-test with effect size and normality tests is presented in Figure 16.



Paired Samples T-Test

Paired Samples T-Test

			statistic	df	p	Cohen's d
pretest	posttest	Student's t	36.6	39.0	< .001	5.79

Tests of Normality

			statistic	p
pretest	posttest	Shapiro-Wilk	0.960	0.169
		Kolmogorov-Smirnov	0.0920	0.856
		Anderson-Darling	0.513	0.183

Descriptives

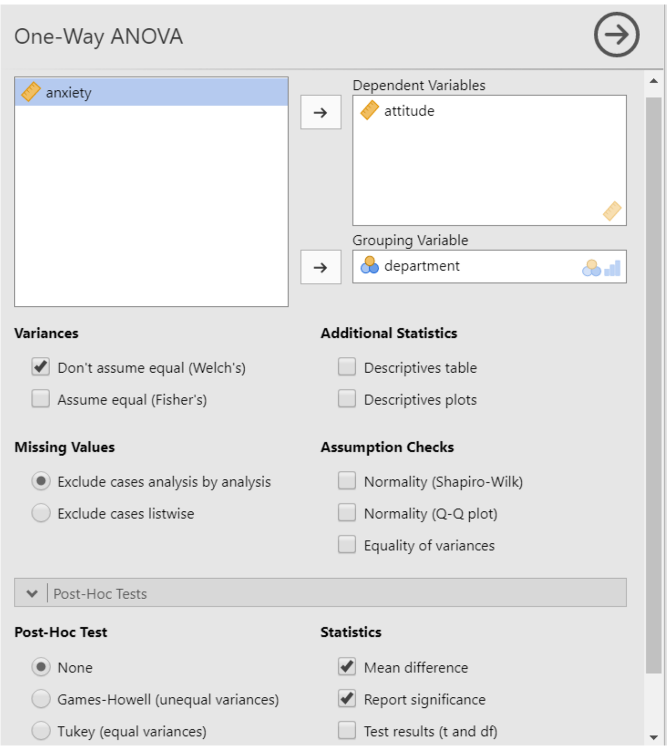
	N	Mean	Median	SD	SE
pretest	40	143.0	140.3	9.16	1.448
posttest	40	76.6	77.1	4.78	0.755

Figure 16. Paired Samples t-test Output

As can be seen from the output, the assumption tests and paired samples t-test results are presented in the same section.

2.6.4. *One Way ANOVA*

When comparing the means of more than two groups, one-way ANOVA is used (Kline, 2009). In Jamovi, this is done by selecting the *One Way ANOVA* path under the *ANOVA* menu and results in a menu such as the one presented in Figure 17.


One-Way ANOVA

anxiety

Dependent Variables: attitude

Grouping Variable: department

Variances
- ☑ Don't assume equal (Welch's)
- □ Assume equal (Fisher's)

Additional Statistics
- □ Descriptives table
- □ Descriptives plots

Missing Values
- ● Exclude cases analysis by analysis
- ○ Exclude cases listwise

Assumption Checks
- □ Normality (Shapiro-Wilk)
- □ Normality (Q-Q plot)
- □ Equality of variances

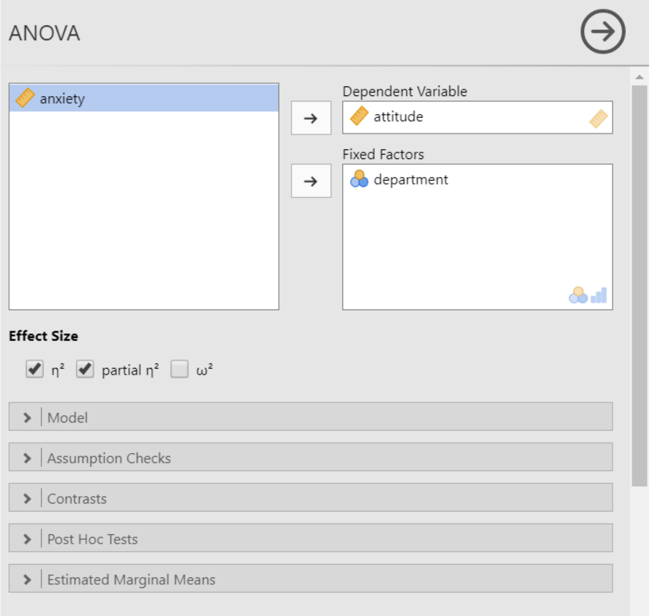
Post-Hoc Tests

Post-Hoc Test
- ● None
- ○ Games-Howell (unequal variances)
- ○ Tukey (equal variances)

Statistics
- ☑ Mean difference
- ☑ Report significance
- □ Test results (t and df)

Figure 17. One Way ANOVA Menu

Descriptive statistics, assumption tests, and post-hoc tests are accessible through the One Way ANOVA menu in Jamovi. However, as may be noticed, post-hoc tests such as effect size and Bonferroni are not present in this section. Options for these exist under the *ANOVA* route of the *ANOVA* menu in Jamovi. This menu tree is presented in Figure 18.


ANOVA

anxiety

Dependent Variable: attitude

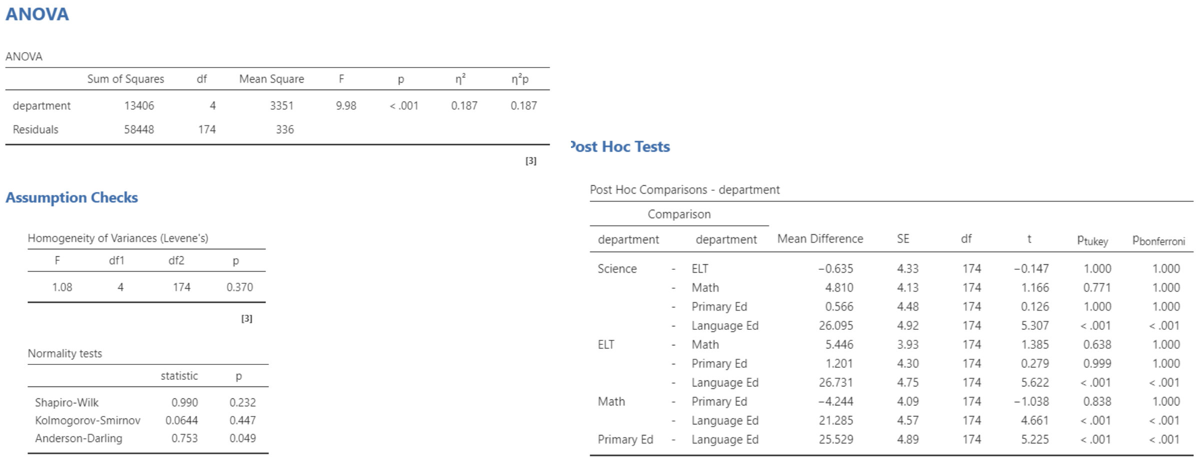
Fixed Factors: department

Effect Size

☑ η² ☑ partial η² □ ω²

Figure 18. ANOVA Menu

The results of the analysis conducted using the options under the ANOVA menu are presented in Figure 19. During these tests, the η^2 and partial η^2 effect size calculations were also selected, while the Bonferroni post-hoc test was selected to determine variance homogenity and normality.



ANOVA

ANOVA

	Sum of Squares	df	Mean Square	F	p	η^2	η^2p
department	13406	4	3351	9.98	< .001	0.187	0.187
Residuals	58448	174	336				

[3]

Assumption Checks

Homogeneity of Variances (Levene's)

F	df1	df2	p
1.08	4	174	0.370

[3]

Normality tests

	statistic	p
Shapiro-Wilk	0.990	0.232
Kolmogorov-Smirnov	0.0644	0.447
Anderson-Darling	0.753	0.049

ᵓost Hoc Tests

Post Hoc Comparisons - department

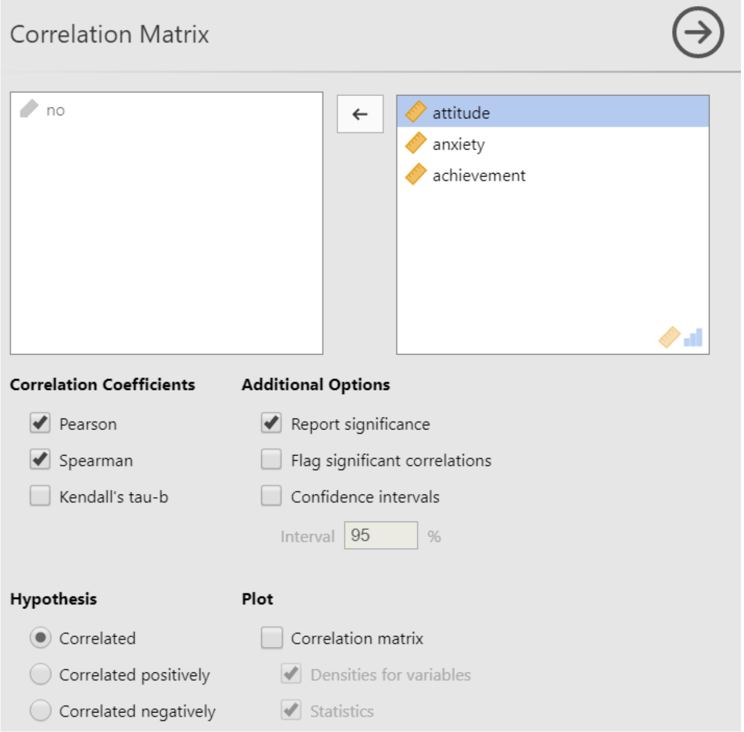
Comparison								
department		department	Mean Difference	SE	df	t	p_{tukey}	$p_{bonferroni}$
Science	-	ELT	−0.635	4.33	174	−0.147	1.000	1.000
	-	Math	4.810	4.13	174	1.166	0.771	1.000
	-	Primary Ed	0.566	4.48	174	0.126	1.000	1.000
	-	Language Ed	26.095	4.92	174	5.307	< .001	< .001
ELT	-	Math	5.446	3.93	174	1.385	0.638	1.000
	-	Primary Ed	1.201	4.30	174	0.279	0.999	1.000
	-	Language Ed	26.731	4.75	174	5.622	< .001	< .001
Math	-	Primary Ed	−4.244	4.09	174	−1.038	0.838	1.000
	-	Language Ed	21.285	4.57	174	4.661	< .001	< .001
Primary Ed	-	Language Ed	25.529	4.89	174	5.225	< .001	< .001

Figure 19. *ANOVA* Output

As can be seen, the ANOVA results, effect sizes, assumption tests and post-hoc test results are presented in the same section.

2.6.5. *Correlation*

When the correlation coefficient is to be calculated to provide information regarding the direction and size of a relationship between two variables (Warner, 2008), the *Regression* menu in Jamovi is chosen and the *Correlation Matrix* is selected. This is followed by a menu presented to users as portrayed in Figure 20.



Correlation Matrix

no

attitude
anxiety
achievement

Correlation Coefficients
☑ Pearson
☑ Spearman
☐ Kendall's tau-b

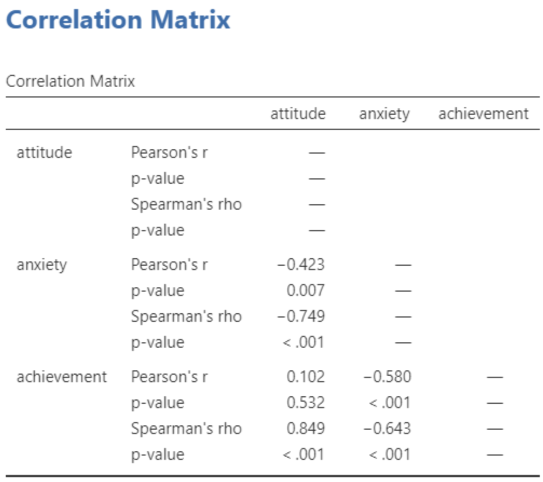
Additional Options
☑ Report significance
☐ Flag significant correlations
☐ Confidence intervals
Interval 95 %

Hypothesis
● Correlated
○ Correlated positively
○ Correlated negatively

Plot
☐ Correlation matrix
☑ Densities for variables
☑ Statistics

Figure 20. Correlation Menu

Using this menu, the variables for which a correlation coefficient is to be calculated are selected, and either one, several, or all of the options for Pearson product-moment correlation coefficient, Spearman rank correlation coefficient, or Kendall's tau-b coefficient are selected. The researcher selects the appropriate options of meaningful coefficients, confidence intervals, graphs for the correlation matrix, and the one-way or two-way nature of the hypotheses being tested. An example for the correlation test output is presented in Figure 21.



Correlation Matrix

Correlation Matrix

		attitude	anxiety	achievement
attitude	Pearson's r	—		
	p-value	—		
	Spearman's rho	—		
	p-value	—		
anxiety	Pearson's r	−0.423	—	
	p-value	0.007	—	
	Spearman's rho	−0.749	—	
	p-value	< .001	—	
achievement	Pearson's r	0.102	−0.580	—
	p-value	0.532	< .001	—
	Spearman's rho	0.849	−0.643	—
	p-value	< .001	< .001	—

Figure 21. Correlation Output

As shown, a correlation matrix was created from the correlation coefficients between the three variables. Additionally, two types of correlation coefficients were reported with their p values as both Pearson and Spearman were selected.

2.6.6. *Linear Regression*

When one or more predictor variable and one dependent variable is to be estimated, linear regression is utilized (Navarro & Foxcroft, 2019). To execute linear regression in Jamovi, *Linear Regression* is selected under the *Regression* menu as shown in Figure 22.

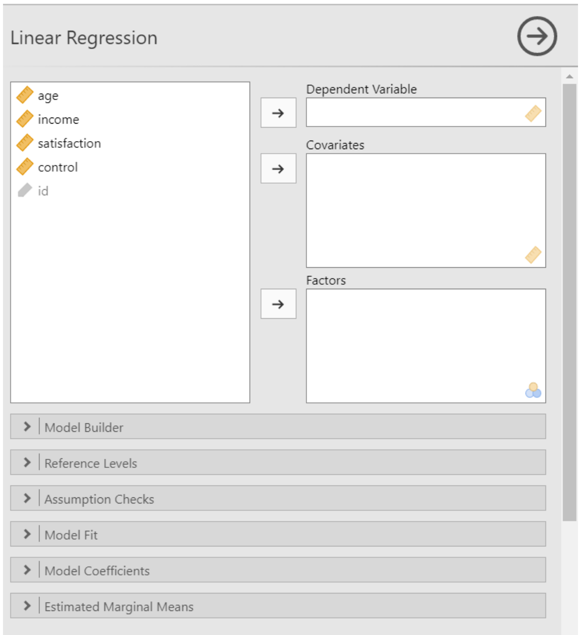
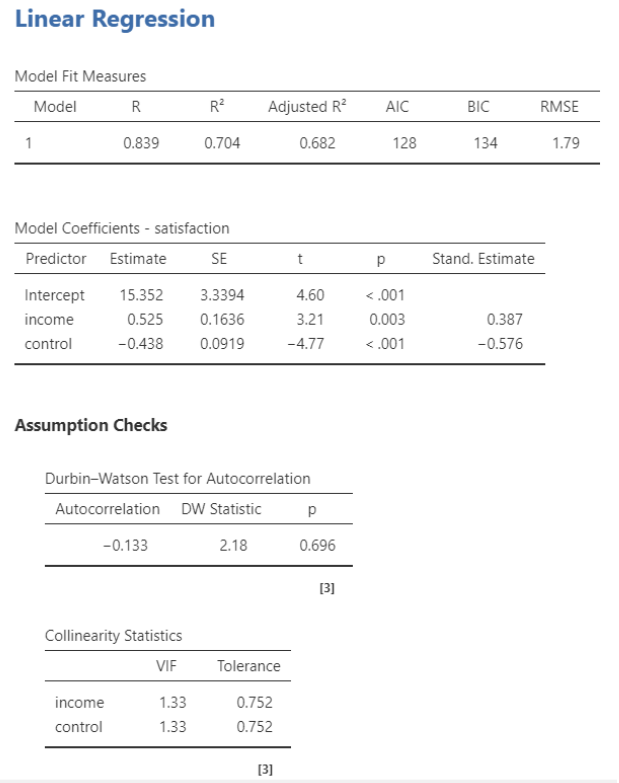


Figure 22. Regression Menu

In this menu, predictor and dependent variables may be defined, and from the assumption control section autocorrelation tests, collinearity statistics, normality tests, and residual plots may be requested. In addition to the R, R^2, and adjusted R^2 values to determine the fit of the regression model, RMSE, F test, AIC and BIC values may also be calculated. Furthermore, standardized coefficients may be retrieved from the Model Coefficients section. The output of a simple multiple regression analysis is provided in Figure 23.



Linear Regression

Model Fit Measures

Model	R	R^2	Adjusted R^2	AIC	BIC	RMSE
1	0.839	0.704	0.682	128	134	1.79

Model Coefficients - satisfaction

Predictor	Estimate	SE	t	p	Stand. Estimate
Intercept	15.352	3.3394	4.60	< .001	
income	0.525	0.1636	3.21	0.003	0.387
control	−0.438	0.0919	−4.77	< .001	−0.576

Assumption Checks

Durbin–Watson Test for Autocorrelation

Autocorrelation	DW Statistic	p
−0.133	2.18	0.696

[3]

Collinearity Statistics

	VIF	Tolerance
income	1.33	0.752
control	1.33	0.752

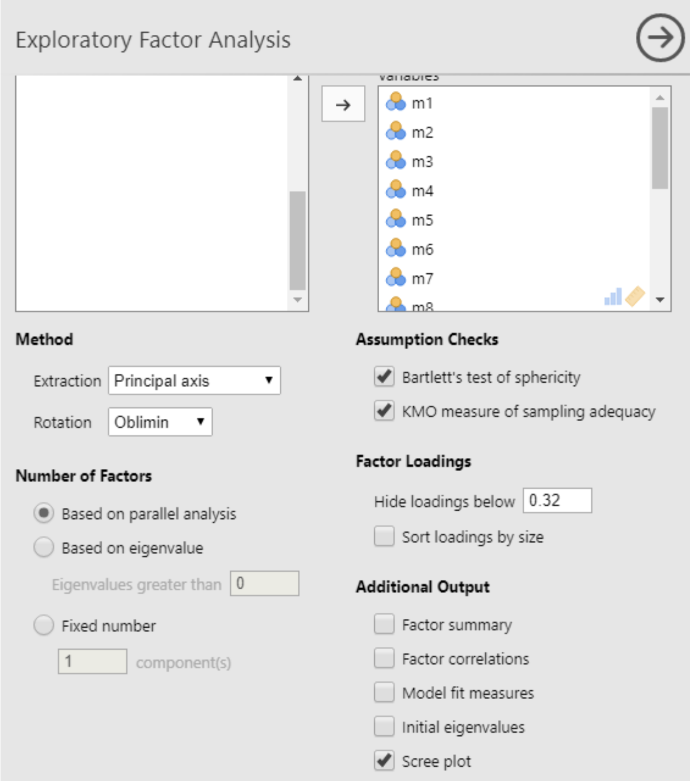
[3]

Figure 23. Linear Regression Output

As can be seen from the output, the Durbin-Watson statistic, VIF, and Tolerance values along with the coefficients of the variables in the model are calculated. Additionally, if the F Test is selected in the *Model Fit* section, the meaningfulness of the model as a whole is tested.

2.6.7. *Exploratory Factor Analysis*

In order to reveal the findings of the construct validity of a measurement tool, EFA is used when there is no strong a priori regarding the components of the structure intended to measure (Henson & Roberts, 2006). PCA is also used for similar purposes to EFA when there is no strong a priori regarding the structure and it is thought to be equivalent to EFA; however, the mathematics it uses makes simple mathematical groupings instead of revealing the latent variable in behavioural sciences, and therefore, it is thought to give biased results (Fokkema & Grieff, 2017). The screen of an EFA in practice is presented in Figure 24.



Exploratory Factor Analysis

Variables
m1
m2
m3
m4
m5
m6
m7
m8

Method

Extraction Principal axis

Rotation Oblimin

Number of Factors

● Based on parallel analysis

○ Based on eigenvalue

Eigenvalues greater than 0

○ Fixed number

1 component(s)

Assumption Checks

☑ Bartlett's test of sphericity

☑ KMO measure of sampling adequacy

Factor Loadings

Hide loadings below 0.32

□ Sort loadings by size

Additional Output

□ Factor summary

□ Factor correlations

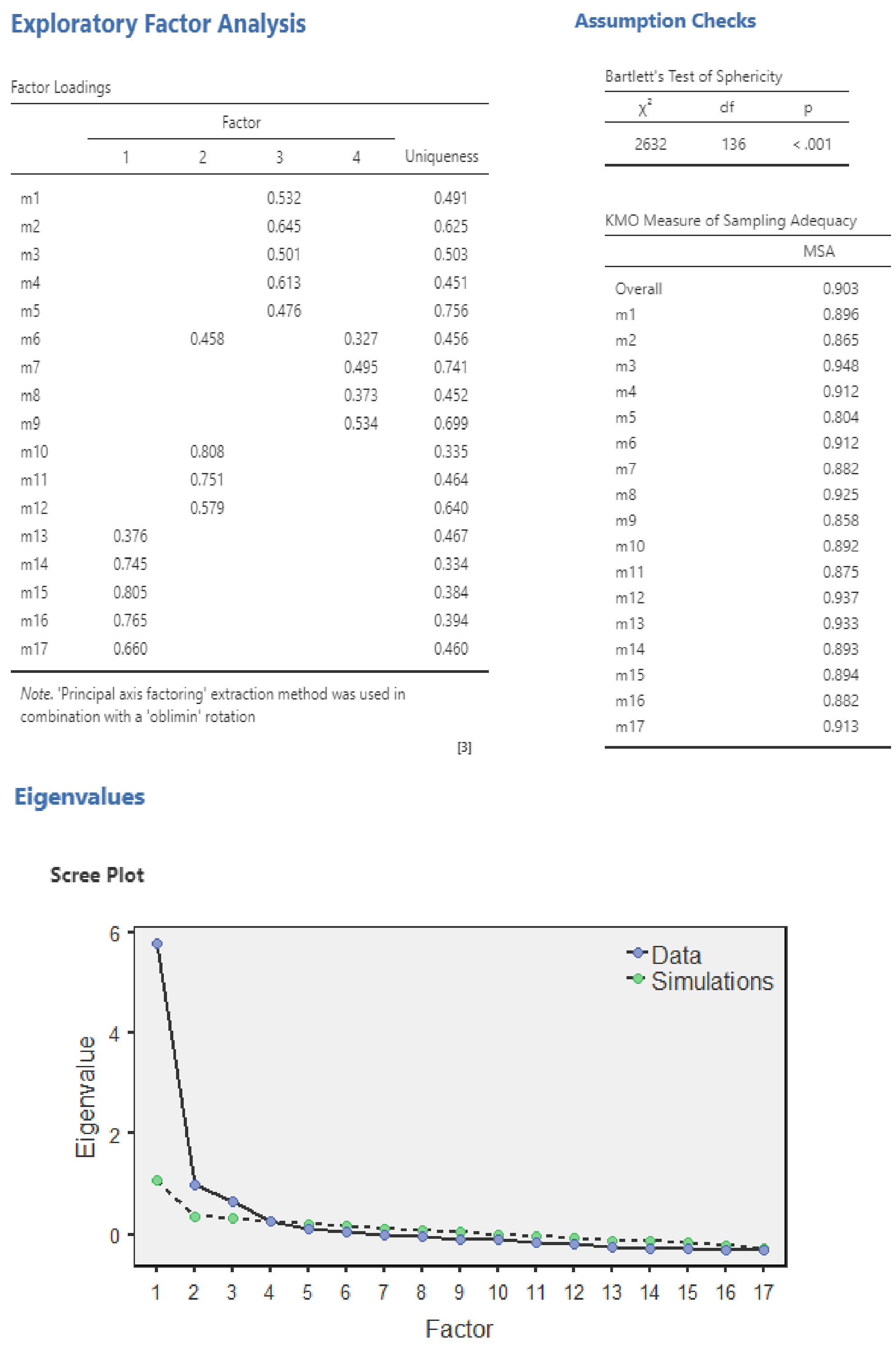
□ Model fit measures

□ Initial eigenvalues

☑ Scree plot

Figure 24. Exploratory Factor Analysis Menu

Firstly, all of the items required to be within the scope of the analysis are entered into the box on the right. From the *Method* heading in the menu, the *Extraction* section is used to select Exploratory Factor Analysis. The *Rotation* section right beneath contains the oblique or orthogonal rotation methods to be used in creating the factor structure. The *Number of Factors* section refers to the method to be used in determining the number of factors and contains the options for *Eigenvalue* or *Parallel Analysis* (Horn, 1965). Also in this section, the option to limit the number of dimensions based on an a priori is presented. On the top right of the menu, in the section regarding the testing of the assumptions of the analysis, the options for *Bartlett's test* and *KMO* which is an analysis regarding the factorizability of items are selectable. In the *Factor Loadings* section right beneath lies a section in which the lowest factor load to be reported can be determined. In the *Additional Output* section of the analysis, options such as *Scree plot*, and *Model fit measures* to obtain fit indices similar to interdimensional correlation or structural equation modeling applications are presented. An output exemplifiying this practice is shown in Figure 25.



Exploratory Factor Analysis

Factor Loadings

	Factor				
	1	2	3	4	Uniqueness
m1			0.532		0.491
m2			0.645		0.625
m3			0.501		0.503
m4			0.613		0.451
m5			0.476		0.756
m6		0.458		0.327	0.456
m7				0.495	0.741
m8				0.373	0.452
m9				0.534	0.699
m10		0.808			0.335
m11		0.751			0.464
m12		0.579			0.640
m13	0.376				0.467
m14	0.745				0.334
m15	0.805				0.384
m16	0.765				0.394
m17	0.660				0.460

Note. 'Principal axis factoring' extraction method was used in combination with a 'oblimin' rotation

[3]

Assumption Checks

Bartlett's Test of Sphericity

χ^2	df	p
2632	136	< .001

KMO Measure of Sampling Adequacy

	MSA
Overall	0.903
m1	0.896
m2	0.865
m3	0.948
m4	0.912
m5	0.804
m6	0.912
m7	0.882
m8	0.925
m9	0.858
m10	0.892
m11	0.875
m12	0.937
m13	0.933
m14	0.893
m15	0.894
m16	0.882
m17	0.913

Eigenvalues

Scree Plot

Figure 25. Exploratory Factor Analysis Output

2.6.8. *Confirmatory Factor Analysis*

CFA is used to determine whether or not the predicted structure is present in the dataset at hand in cases where a strong a priori regarding the structure is apparent to the researcher (Brown, 2015). Therefore, it may be stated that CFA is used when a new or previously developed scale for a structure with theoretically sound dimensions is to be used on a different sample or adapted to a different culture. An example for a 5-item single dimension scale in a CFA application is presented in Figure 26.

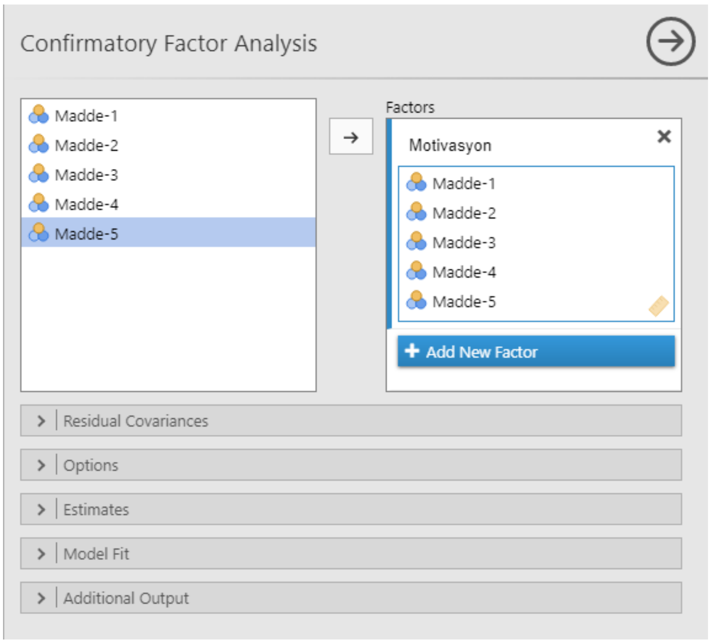
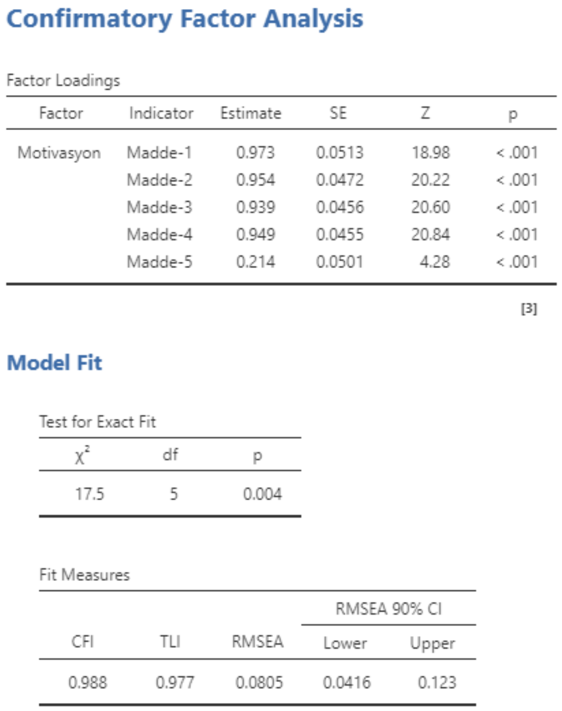


Figure 26. Confirmatory Factor Analysis Menu

In the top of the screen seen in Figure 26, the section in which the dimensions for the items to be analyzed are defined can be seen. Here, the latent variable is given a name and the items related to it are moved to the box below it. If there is more than one dimension in the scale, the "*Add new factor*" command is clicked and in the new box that opens, the same process for the first latent variable is repeated for the new factor.

The *Residual Covariances* tab is used to define the covariance between the residual variances of the observed variable to improve the model-data fit when necessary. However, to conduct this operation, the *Modification indices* command under the *Additional Output* tab at the bottom must be selected to obtain the suggested modifications. Under the *Options* heading, the functions regarding how to deal with missing values are present under *Missing Values Method* while the latent variable scale assignment which needs to be performed before analysis in structural equation modeling techniques can be done through the *Constraints* function. In the Estimates tab lie the *Results* and *Statistics* sections. The z and p values and confidence intervals for the estimated factor loads may be requested from here. In the *Model Fit* tab, the reported fit indices regarding the model-data fit in structural equation modeling are presented. While the values for χ^2, CFI, TLI, and RMSEA are provided as default, the values for SRMR, AIC, and BIC may also be selected. Lastly, the path diagram for the visualization of the model may be requested from the *Additional Output* tab mentioned previously on the subject of modification indices. The output obtained for this model is presented in Figure 27.



Confirmatory Factor Analysis

Factor Loadings

Factor	Indicator	Estimate	SE	Z	p
Motivasyon	Madde-1	0.973	0.0513	18.98	< .001
	Madde-2	0.954	0.0472	20.22	< .001
	Madde-3	0.939	0.0456	20.60	< .001
	Madde-4	0.949	0.0455	20.84	< .001
	Madde-5	0.214	0.0501	4.28	< .001

[3]

Model Fit

Test for Exact Fit

χ^2	df	p
17.5	5	0.004

Fit Measures

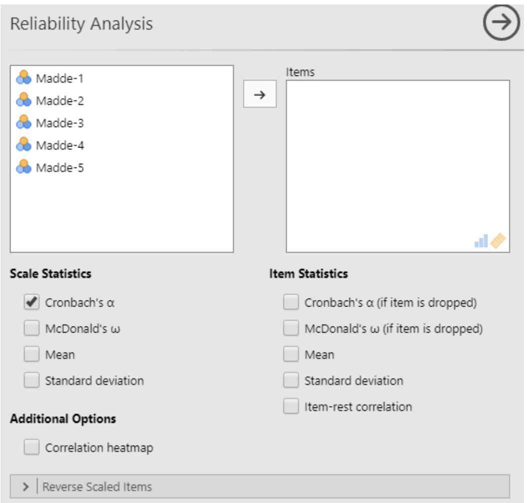
			RMSEA 90% CI	
CFI	TLI	RMSEA	Lower	Upper
0.988	0.977	0.0805	0.0416	0.123

Figure 27. Confirmatory Factor Analysis Output

2.6.9. *Reliability*

Reliability, in a broad sense, refers to the degree to which the measurement tool is free of random faults (Baykul, 2000). While reliability also has other definitions and meanings, its most frequent use is through the Cronbach Alpha (α) coefficient as an indicator of internal consistency due to its applicability and interpretability (Padilla & Divers, 2016). In addition to Cronbach's α, Jamovi also allows for the reporting of McDonald's Omega (ω) coefficient, which has lately been included in recent research as "composite reliability", as an indicator of reliability in the sense of internal consistency.

The Reliability tab under the Factor option in Jamovi results in the screen shown in Figure 28.



Reliability Analysis

Madde-1
Madde-2
Madde-3
Madde-4
Madde-5

Items

Scale Statistics

☑ Cronbach's α
□ McDonald's ω
□ Mean
□ Standard deviation

Item Statistics

□ Cronbach's α (if item is dropped)
□ McDonald's ω (if item is dropped)
□ Mean
□ Standard deviation
□ Item-rest correlation

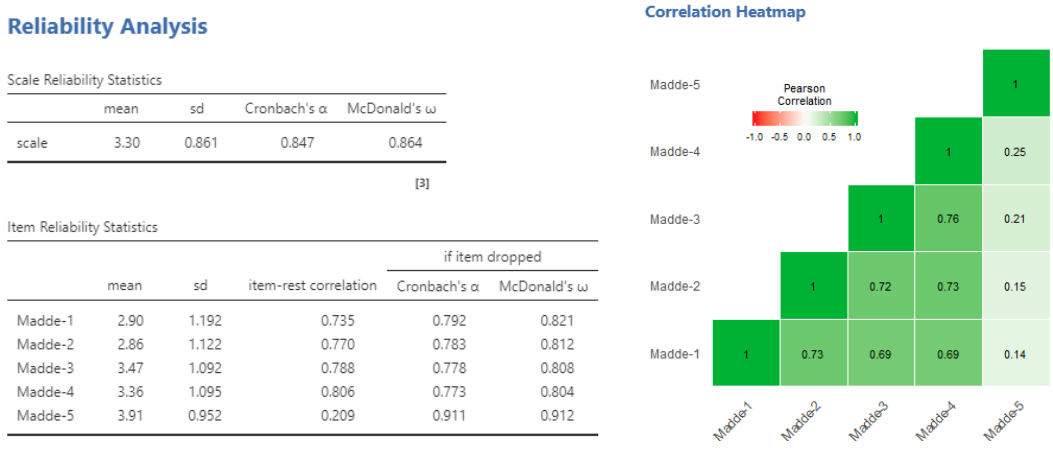
Additional Options

□ Correlation heatmap

> Reverse Scaled Items

Figure 28. Reliability Analysis Screen

On the screen portrayed in Figure 28, first the items to be used in the analysis are selected and moved to the box on the right. In the *Scale Statistics* section, the options to request Cronbach's α and McDonald's ω coefficients along with the arithmetic mean and standard deviation of the scale are presented. Under the *Item Statistics* section are values regarding items. Here, the change in the reliability coefficient, arithmetic means of the items, and standard deviation in the event that any one item is removed from the scale along with the correlation of each individual item with the test as a whole may be requested. In the *Additional Options* section, the correlation of items with each other is presented with a heat map. The *Reverse Scaled Items* tab at the bottom aids in determining items with inverted meanings to speed up processes. The outputs of Reliability analyses are shown in Figure 29.



Reliability Analysis

Scale Reliability Statistics

	mean	sd	Cronbach's α	McDonald's ω
scale	3.30	0.861	0.847	0.864

[3]

Item Reliability Statistics

				if item dropped	
	mean	sd	item-rest correlation	Cronbach's α	McDonald's ω
Madde-1	2.90	1.192	0.735	0.792	0.821
Madde-2	2.86	1.122	0.770	0.783	0.812
Madde-3	3.47	1.092	0.788	0.778	0.808
Madde-4	3.36	1.095	0.806	0.773	0.804
Madde-5	3.91	0.952	0.209	0.911	0.912

Correlation Heatmap

Figure 29. Reliability Analysis Output

2.6. Extending Features through Modules

Beyond the basic analyses Jamovi provides, other analyses may be conducted through modules. The *Modules* button can be used to see, review, and load modules developed for Jamovi (Figure 30).

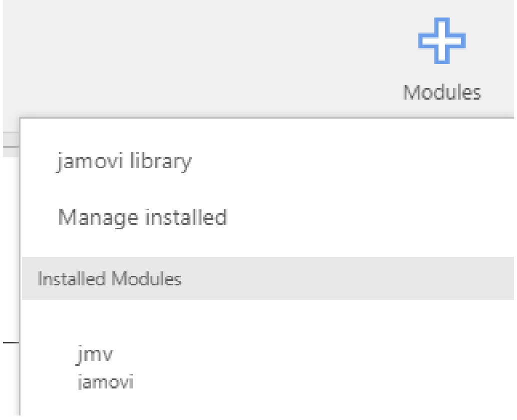
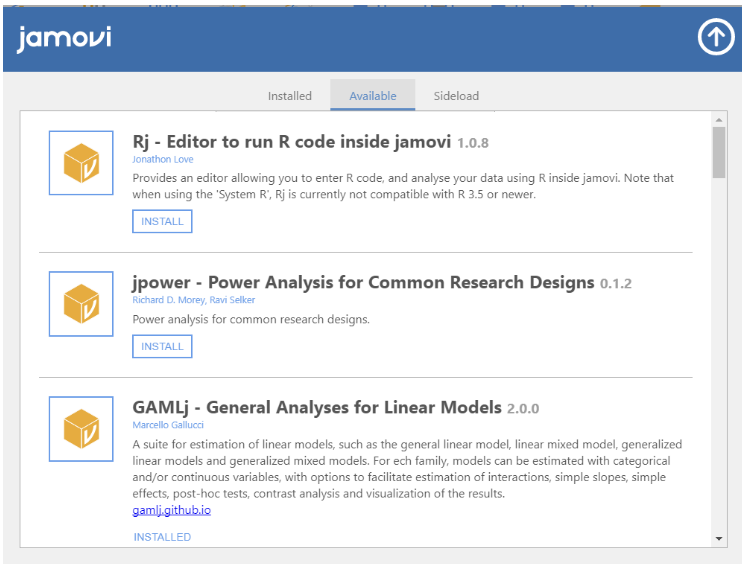


Figure 30. Modules Menu

By following the *Modules* and *Jamovi library* path, all of the modules that can be loaded can be listed as shown in Figure 31.

As of the date this article was written, the Jamovi module library was home to many modules such as, *Rj* which allows R code to be run in Jamovi, *jpower* which allows for Power analysis, *GAMLj* for linear models, *jsq* for Bayesian methods, and *MAJOR* for meta-analysis. Jamovi modules continue to expand with the contributions of developers.



Rj - Editor to run R code inside jamovi 1.0.8
Jonathon Love
Provides an editor allowing you to enter R code, and analyse your data using R inside jamovi. Note that when using the 'System R', Rj is currently not compatible with R 3.5 or newer.
INSTALL

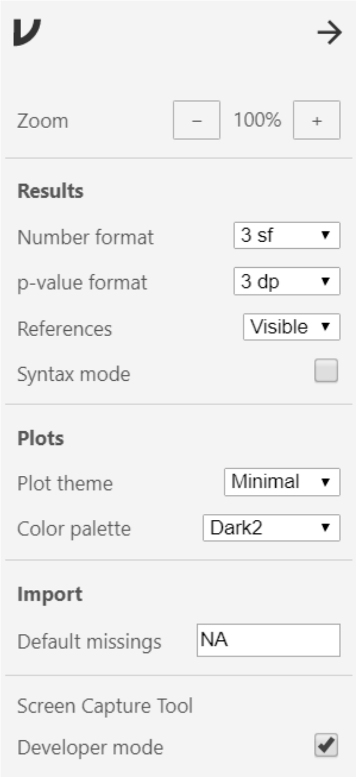
jpower - Power Analysis for Common Research Designs 0.1.2
Richard D. Morey, Ravi Selker
Power analysis for common research designs.
INSTALL

GAMLj - General Analyses for Linear Models 2.0.0
Marcello Gallucci
A suite for estimation of linear models, such as the general linear model, linear mixed model, generalized linear models and generalized mixed models. For ech family, models can be estimated with categorical and/or continuous variables, with options to facilitate estimation of interactions, simple slopes, simple effects, post-hoc tests, contrast analysis and visualization of the results.
gamlj.github.io
INSTALLED

Figure 31. Jamovi Modules

2.7. Options

Jamovi has certain basic options regarding theming and number portrayal. Clicking on the three dots on the top right corner provides access to these settings. The settings menu that appears is shown in Figure 32.



Zoom − 100% +

Results
Number format: 3 sf
p-value format: 3 dp
References: Visible
Syntax mode ☐

Plots
Plot theme: Minimal
Color palette: Dark2

Import
Default missings: NA

Screen Capture Tool
Developer mode ☑

Figure 32. Jamovi Options

From this menu, basic options such as number and p value format, showing or hiding references, and graphics themes can be set.

3. AVAILABILITY

To download Jamovi, visit https://www.jamovi.org, and to access the source code of the jamovi project visit the github page at https://github.com/jamovi/. Additionally, a detailed explanation of Jamovi prepared by Navarro and Foxcroft (2019) is available as a free e-book from https://www.learnstatswithjamovi.com/.

Conflicts of Interest

The authors declare no conflict of interest.

ORCID

Murat Doğan Şahin https://orcid.org/0000-0002-2174-8443
Eren Can Aybek https://orcid.org/0000-0003-3040-2337

4. REFERENCES

Amazon (2019). *Search Results. https://www.amazon.com/s?k=statistics+for+social+science* Accessed on 20 September 2019.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme [Measurement and Evaluation in Education]*. Ankara: ÖSYM Yayınları.

Brown, T.A. (2015). *Confirmatory factor analysis for applied research.* New York: The Guilford Press.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: SAGE Pub.

Fokkema, M. & Grieff, S. (2017). How performing PCA and CFA on the same data equals trouble. *European Journal of Psychological Assessment*, *33*(6), 399–402.

GNU (2019). *What is GNU?* https://www.gnu.org/ Accessed on 28 September 2019.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179-85.

Henson, H. K. & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*(3). 393-416.

JASP Team (2019). *JASP* (Version 0.11.1) [Computer software].

Kline, R. B. (2009). *Becoming a behavioral science researcher: A guide to producing research that matters*. New York: Guilford Press.

Muenchen, R.A. (2019). *The Popularity of Data Science Software*. http://r4stats.com/articles/popularity/. Accessed on 20 September 2019.

Navarro, D.J. & Foxcroft, D.R. (2019). *Learning statistics with jamovi: A tutorial for psychology students and other beginners*. (Version 0.70). doi: 10.24384/hgc3-7p15 [Available from http://learnstatswithjamovi.com]

Padilla, M. A & Divers, J. (2016). A Comparison of composite reliability estimators: Coefficient omega confidence intervals in the current literature. *Educational and Psychological Measurement*, *76*(3) 436–453

R Core Team (2019a). *What is R.* https://www.r-project.org/about.html Accessed on 25 September 2019.

R Core Team (2019b). *Contributed Packages. https://cloud.r-project.org/web/packages* Accessed on 01 October 2019.

RKWard (2019). *About RKWard.* https://rkward.kde.org/About.html Accessed on 05 October 2019.

RStudio (2019). *RStudio.* https://rstudio.com/products/rstudio/ Accessed on 05 October 2019.

The jamovi project (2019a). *Jamovi* (Version 1.0) [Computer Software].

The jamovi project (2019b). *About.* https://www.jamovi.org/about.html Accessed on 06 October 2019.

Warner, R.M. (2008). *Applied statistics: From bivariate through multivariate techniques.* USA: SAGE Pub.

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 693–705
https://dx.doi.org/10.21449/ijate.576249

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

Classroom Response Systems as a Formative Assessment Tool: Investigation into Students' Perceived Usefulness and Behavioural Intention

Muhittin Şahin [1,*]

[1]Department of Computer Education and Instructional Technology, Ege University, 35040, Izmir, Turkey

ARTICLE HISTORY

Received: 15 April 2019

Revised: 26 July 2019

Accepted: 10 November 2019

KEYWORDS

Formative assessment,
Classroom response systems,
Perceived usefulness,
Behavioral intention.

Abstract: Assessments are conducted to determine the effectiveness of learning. One type of these assessments are formative assessment, which aims to fill the gap between the learner's present situation and the desired situation by giving feedback to learners. For this purpose, Classroom Response Systems can be used in large groups. Paper-based tests, Kahoot, Quizizz, and, Plickers were used for formative assessment. Multiple-choice tests can be created for students with these applications. Students can connect to Kahoot and Quizizz applications via any computer, tablet or mobile phones with an internet connection and answer the questions in the test. For the Plickers application, the questions are displayed by the instructor in an area that all students can see. Students indicate their responses by lifting their paper which has QR code. The instructor scans these QR codes with the help of a mobile device and the students' answers are seen directly. In this context, the perceived usefulness and behavioral intention of the students to use different classroom response systems were investigated. The research was conducted with freshman students at a state university and continued for four weeks. Different applications were presented to the students as like that a paper-based test, Kahoot, Quizizz, and Plickers. When the findings were examined, it was found that students noted Kahoot, Quizizz, and Plickers applications were statistically more useful than the paper-based test. Based on these results, it can be said that students prefer to use technology-supported classroom response systems instead of the paper-based test.

1. INTRODUCTION

Assessment is a basic component of effective learning (Bransford, Brown, & Cocking, 2000). there were two instructional assessments as formative and summative in the past. Nowadays assessment was reclassified as Assessment of Learning (AoL), Assessment for Learning (AfL) and Assessment as Learning (AaL). While AoL corresponds to the summative assessment (SA), the formative assessment (FA) is divided into two subgroups as AfL and AaL. FA aims to reform the learning (Bayrak & Yurdugül, 2016). One of the foundation of FA is to fill the gap between the desired purpose and the existing situation of the learner (Black & Wiliam, 1998).

CONTACT: Muhittin Şahin ✉ muhittin.sahin@ege.edu.tr Department of Computer Education and Instructional Technology, Ege University, 35040, Izmir, TURKEY

ISSN-e: 2148-7456 /© IJATE 2019

Formative assessment is a process in which various tools and strategies are used to determine what the student knows, to improve learning and to plan future teaching (Pinchok & Brandt, 2009). Quality feedback is the key feature of FA (Black & Wiliam, 1998). Feedback contains important information about students' learning. This information is important for the instructor (AfL) to restructure teaching, and important for students (AaL) to increase their learning awareness and improvelearning experiences.

However, due to the crowded classrooms, the instructors have some problems in both preparing questions and giving feedback about missing concepts (Bayrak & Yurdugül, 2016). At this point, classroom response systems (CRS) which can be applied to large masses, helped to solve this problem. It is possible to give immediate and quality feedback to the students and instructors through CRS (Kay, 2009; Lucke, Keyssner, & Dunn, 2013; Fuller & Dawson, 2017). Within the scope of this research, CRSs are discussed in the context of FA because by nature these systems are used for formative assessment (Beatty, Gerace, Leonard, & Dufresne, 2006).

Determination of perceived usefulness and behavioral intention are important for using the CRSs. These concepts are explained by acceptance and adoption theories and models. Perceived usefulness, ease of use, facilitating conditions and social influence factors play an important role in acceptance and adoption theories and models (Usluel & Mazman, 2010). And also, the level of perceived usefulness by the individual about innovation is directly related to use (Rogers, 2003). Therefore, in the context of this research, perceived usefulness and behavioral intention were included. Usluel and Mazman (2010) define perceived usefulness as the belief that individuals will increase performance by using something new. Behavioral intention is defined as individual readiness of display behavior (Fishbein & Ajzen, 1975). Within the scope of this study, perceived usefulness and behavioral intention about different classroom response systems were investigated.

1.1. Classroom Response System (CRS)

It is seen that CRS has a history of approximately 60 years. The purpose of these systems, which were introduced in the 1960s (called electronic response systems in the 1960s), is to give learners immediate feedback on multiple-choice questions and to inform teachers about the understanding of learners (Judson & Sawada, 2002). It is possible to determine which concepts are missing or misunderstood by the learners through feedback presented to the instructors. It also provides feedback to instructors on what subjects are needed for extra teaching and in which topics students are successful (Bartsch & Murphy, 2011). In addition, the instructors can provide information about the item difficulty and item discrimination, or information about distractors. And also, the systems can provide feedback to learners about individual shortcomings.

CRSs are technologies which used to encourage active learning (Martyn, 2007). Through integrating the CRS to curriculum design, a new communication channel is provided and the classroom interaction between the learner and the instructor can be changed (Siau, Sheng, & Nah, 2006). CRSs are enjoyable and helpful systems that create a communication channel between learners and tutorials in very large classroom environments (Vetterick, Garbe, Dähn, & Cap, 2014). These systems can be used in very large masses or with small groups. CRS has the following features;

- Presentation and ask questions
- Learner response and display
- Data management and analysis (Deal, 2007).

Using CRS has some difficulties and researchers should pay attention to these issues. These difficulties are expressed by Feldman and Capobianco (2003) as follows:

- Create and adaptive appropriate questions
- Create productive classroom environments

The quality of the questions which are prepared in these systems is very important. The learning environment must be increased learner engagement and giving adequate time to students for responding to the questions (Lucke, Keyssner, & Dunn, 2013). Learning environments should be provided with a high level of interaction in order to provide productive classroom environments. In this context, two important characteristics of interaction communication and engagement are encountered (Sims, 2003). In other words, learning environments must be haven some features as increase engagement and encoure communication.

In the literature, it's seen that; CRSs increase learner engagement, participant, and peer interaction (Martyn, 2007; Lucke, Keyssner, & Dunn, 2013; Petto, 2019; Cheng & Wang, 2019; Yılmaz & Karaoğlan Yılmaz, 2019). Another result is that these systems contribute to learning outcomes by increasing their interaction with lectures, tutorials, and other learners (Bartsch & Murphy, 2011; Lucke, Keyssner, & Dunn, 2013; Yang et al., 2019). The literature review study shows that (Kay & LeSage, 2009);

- In general, learners have a positive attitude.
- In the classroom, these systems increased participation, attention, and engagement.
- In the context of learning, it provides interaction and discussion environments, improves learning performance and quality of learning.
- Giving feedback, formative assessment, and compare features are expressed as positive characteristics

Recently, it is possible to say that CRS development and diversity have increased through the improvement of web technologies. As a question type in addition to multiple-choice questions, drag-and-drop, puzzle, and similar applications can also be developed. In addition to integrating the elements of gamification into the systems, more enjoyable and entertaining environments have been started to be developed. In the context of data management and analysis, not only descriptive analysis but also more advanced analysis and presentation of this information have become feasible.

The aim of this study, determining perceived usefulness and behavioral intention levels of using different classroom response systems by students. For this purpose, the following sub-problems were tried to response:

I. Are there any differences in learners' perceived usefulness levels about different CRS?
II. Are there any differences in learners' behavioral intention levels about different CRS?

2. METHOD

In this section research design, participants, data collection tools, data analysis, implementation process and, CRSs used for this study are presented.

2.1. Research Design

This study was structured as survey research. The main purpose of survey research studies is to define the character in a sample (Fraenkel, Wallen, & Hyun, 2011). Within the scope of this research, the perceived usefulness and behavioral intention of using different classroom response systems by students were tried to be determined.

2.2. Participants

The participants of the research are junior students who are assigned Information Technologies course at a public university. Different applications were presented to the students as like that a paper-based test, Kahoot, Quizizz, and Plickers. Respectively, 61, 60, 59 and 54 students participated in these applications. Participants were from three different departments as

Preschool Education, Social Sciences Education, and Guidance and Counseling. The data collection tool was applied to the students at different times after each application. So there are four demographic information about the participants. Detailed information about the participants is presented in the findings section.

2.3. Data Collection Tools

"*Computer Based Assessment Acceptance Model Scale*" was used as data collection tools. The scale includes six factors as "*perceived usefulness, perceived ease of use, computer self-efficacy, social impact, content, and behavioral intention*". But within the scope of this research, is limited to only two factors. The scale developed by Yurdugül and Bayrak (2014) was used in order to determine the characteristics of students' perceived usefulness and behavioral intention. There are three items for perceived usefulness and three items for behavioral intention factors. The scale form was a five-point Likert scale as "Strongly Agree (5)", "Agree (4)", "Undecided (3)", "Disagree (2)" and "Strongly Disagree (1)". The Cronbach Alpha reliability coefficient for the behavioral intention 0,89 and for the perceived usefulness 0,90 was calculated by Yurdugül and Bayrak (2014). Within the scope of this research, Cronbach Alpha reliability coefficients were calculated for each application separately. The obtained reliability coefficients are presented in Table 1.

Table 1. Cronbach Alpha reliability coefficients of the factors

Structure	Paper-based test	Kahoot	Quizizz	Plickers
Behavioral intention	0,93	0,94	0,91	0,94
Perceived usefulness	0,88	0,80	0,83	0,76

According to Table 1, the reliability coefficients are between 0,76 and 0,94 for the scale structures. If these coefficients are greater than 0,70, the results are reliable (Nunnally & Bernstein, 1994; as cited in Yurdugül & Bayrak, 2014). It is possible to say that the factors are reliable.

2.4. Data Analysis

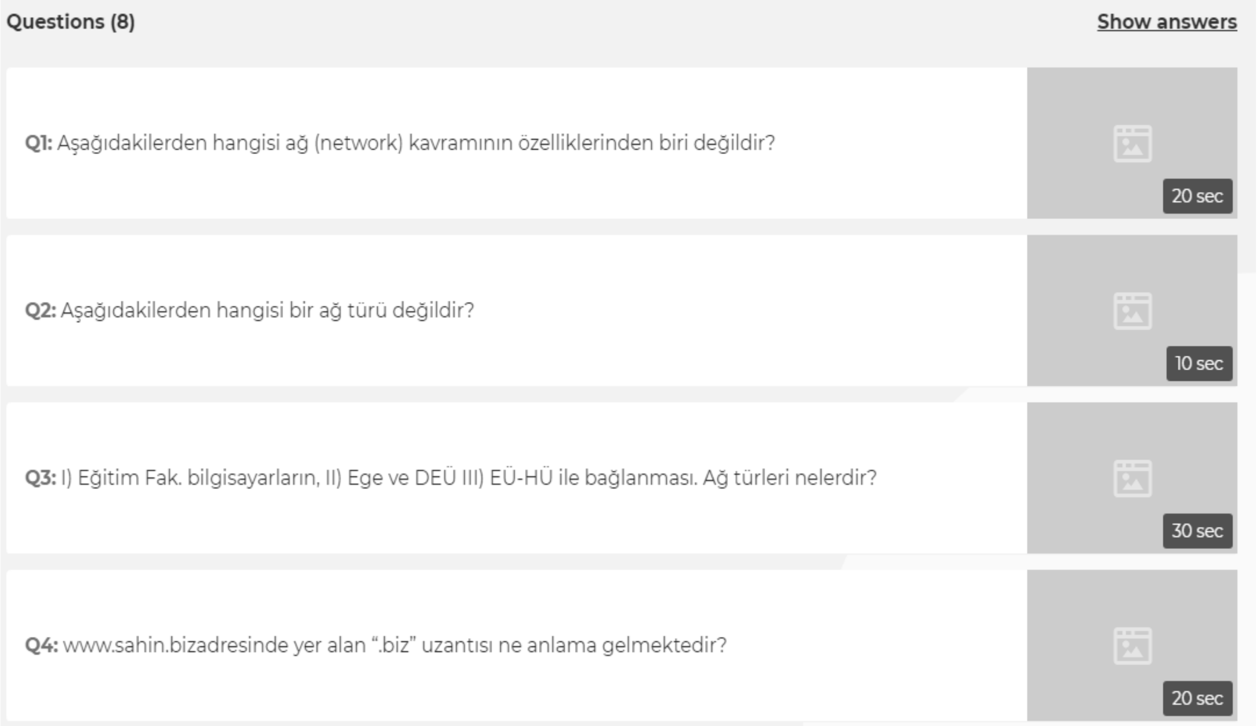
In order to determine differences in the students' perceptions of usefulness and behavioral intention based on the different applications, ANOVA was utilized. Prior to this analysis, assumptions (normal distribution and homogeneity of variances) were tested. For this purpose, firstly, whether there is an extreme value or not was examined. There were no extreme values in the dataset. After this process, in order to ensure the assumptions square transformation was utilized. Then, analyses were done and the findings were interpreted.

2.5. Classroom Response Systems

Within the scope of this research Kahoot, Quizizz and Plickers were chosen and applied as CRS. These CRSs were compared to each other and compared to the paper-based test. In this section, information about the CRSs is given.

2.5.1. *Kahoot*

The Kahoot application is available at https://kahoot.com/. In this application, multiple-choice tests can be created for students. The test is created by the instructor and a game pin number is given to the learners. Students can connect to this application with any computer, tablet or mobile phone with an internet connection and answer the questions in the test. Figure 1 and Figure 2 show screenshots of this application.



Questions (8)

Show answers

Q1: Aşağıdakilerden hangisi ağ (network) kavramının özelliklerinden biri değildir? 20 sec

Q2: Aşağıdakilerden hangisi bir ağ türü değildir? 10 sec

Q3: I) Eğitim Fak. bilgisayarların, II) Ege ve DEÜ III) EÜ-HÜ ile bağlanması. Ağ türleri nelerdir? 30 sec

Q4: www.sahin.bizadresinde yer alan ".biz" uzantısı ne anlama gelmektedir? 20 sec

Figure 1. Kahoot application

Join with the **Kahoot! app** or at **kahoot.it**

with Game PIN:

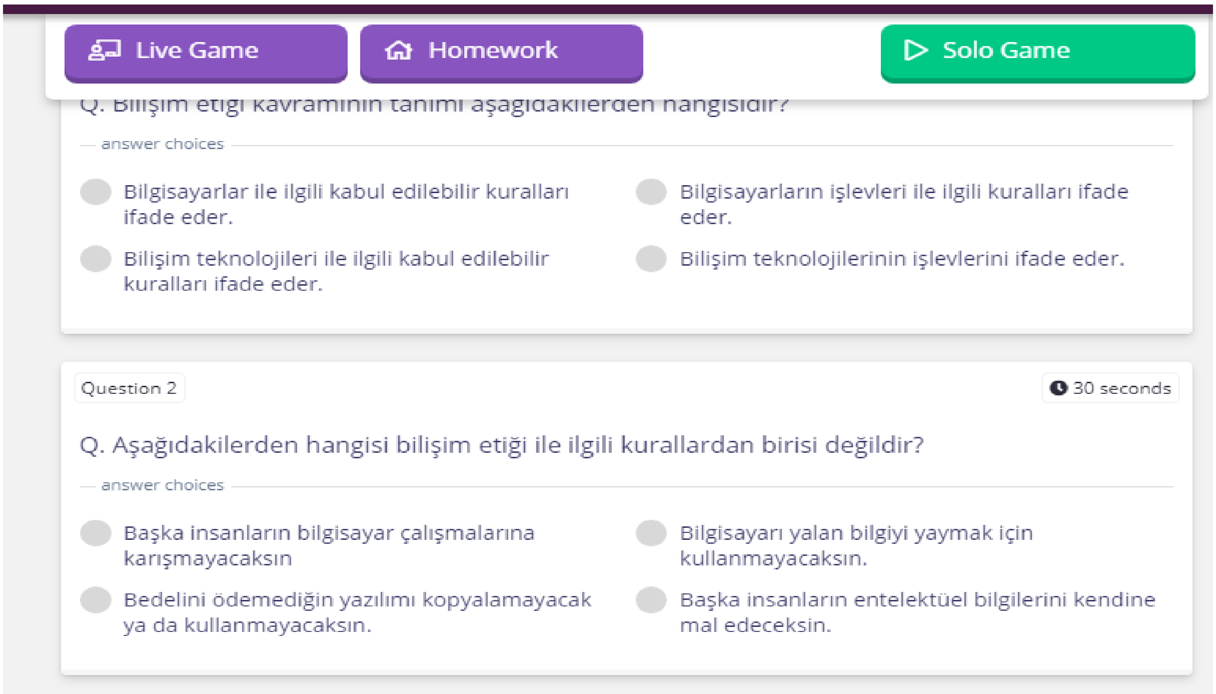
70855

Figure 2. Pin number for Kahoot

As seen in Figure 1, a test consisting of eight questions about the *"Information Technology – Operating Systems"* course was prepared and presented to the students and each question had four choices. Figure 2 shows the pin number for the students who want to join the game. The students entered the pin number at https://kahoot.it/ and after the test was started by the instructor, they answered the questions. There was a different time limit for each question as five, ten, twenty, and thirty seconds. While the Kahoot application was applied, the students saw the questions on the board via a projector and they responed the questions with their computers or mobile devices. Then, the details about the test results were presented to the students. Finally, the details of the application were introduced to the students and students developed a trial test in order to have an experience.

2.5.2. *Quizizz*

The Quizizz application is available at https://quizizz.com/. In this application, multiple-choice tests, puzzle, drag, and drop practice can be created for students. Figure 3 and Figure 4 show screenshots of this application.



Q. Bilişim etiği kavramının tanımı aşağıdakilerden hangisidir?

answer choices

○ Bilgisayarlar ile ilgili kabul edilebilir kuralları ifade eder.

○ Bilgisayarların işlevleri ile ilgili kuralları ifade eder.

○ Bilişim teknolojileri ile ilgili kabul edilebilir kuralları ifade eder.

○ Bilişim teknolojilerinin işlevlerini ifade eder.

Question 2

30 seconds

Q. Aşağıdakilerden hangisi bilişim etiği ile ilgili kurallardan birisi değildir?

answer choices

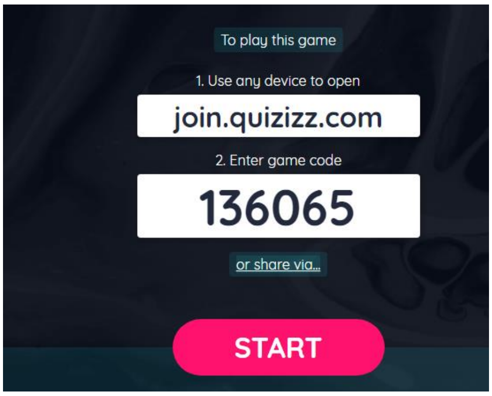
○ Başka insanların bilgisayar çalışmalarına karışmayacaksın

○ Bilgisayarı yalan bilgiyi yaymak için kullanmayacaksın.

○ Bedelini ödemediğin yazılımı kopyalamayacak ya da kullanmayacaksın.

○ Başka insanların entelektüel bilgilerini kendine mal edeceksin.

Figure 3. Quizizz test



To play this game

1. Use any device to open

join.quizizz.com

2. Enter game code

136065

Figure 4. Pin number for Quizizz

As seen in Figure 3, a test consisting of eight questions about the *"Information Technology - Computer Networks"* course was prepared and presented to the students. Figure 4 shows the pin number for the students who want to join the game. In this application, the students can see and response the test questions on their own computer tablet or mobile phone. With the projection device, instant scores and performances of the learners were displayed. And the details about the test results were presented to the students. Then, the details of the application were introduced to students and they developed a trial test in order for them to have experience.

2.5.3. *Plickers*

The Plickers application is available at https://get.plickers.com/. The Plickers application is similar to clickers, which is another CRS. However, each student is not given a clicker to carry out this practice; the students are given a paper-based defined to themselves with QR code. The questions were displayed by the instructor in an area that all students can see. Students indicated their responses by lifting this paper. The instructor read these QR codes with the help of a mobile device and the students' responses were seen directly. Not every student needs a computer, phone or tablet to perform this application.

2.6. Implementation Process

The implementation period continued for four weeks. This study was conducted in *"Information Technology"* course and, about *"Basic Concepts of Information Technology"*, *"Operating Systems"*, *"Computer Networks"* and *"Information Technology Ethics"* subjects. Each week, the course subjects were explained and presented. In the next lesson, the students were tested in accordance with the applications. After the test, the students completed the scale forms online in order to determine the perceived usefulness and behavioral intention. In this study, the tests consisting of eight questions were prepared by the researcher. Totally four tests were developed, one for each subject. The aim of the development of these tests is to give the learners an opportunity toexperience with paper-based, Kahoot, Quizizz and Plickers applications. So the validity and reliability analyses were not performed. The implementation process is presented in Figure 5.

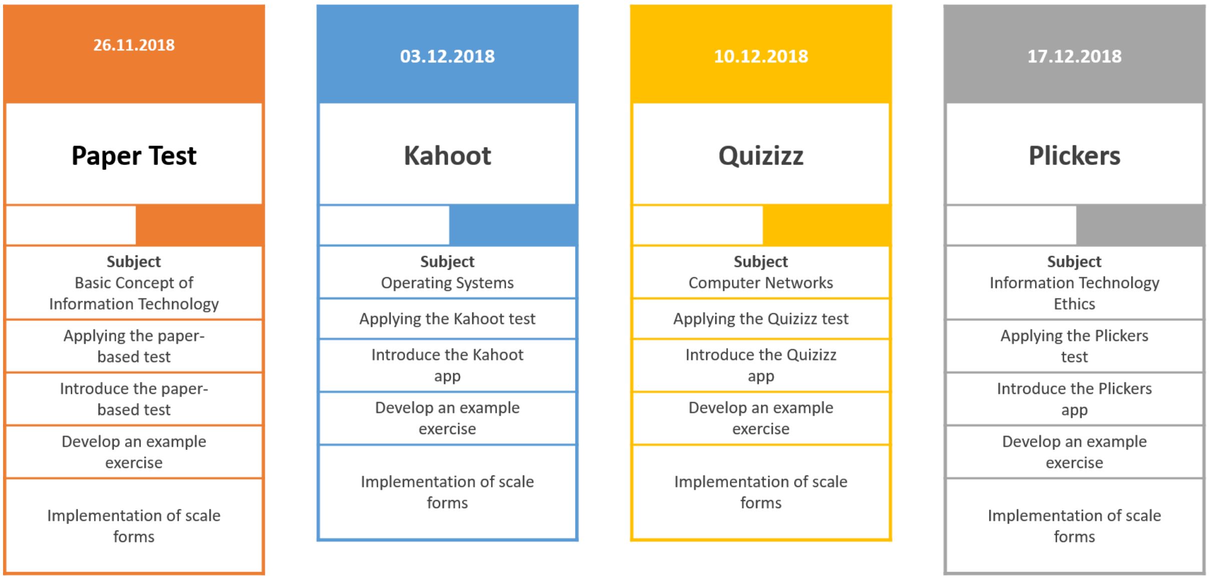

26.11.2018
Paper Test
Subject
Basic Concept of Information Technology
Applying the paper-based test
Introduce the paper-based test
Develop an example exercise
Implementation of scale forms
03.12.2018
Kahoot
Subject
Operating Systems
Applying the Kahoot test
Introduce the Kahoot app
Develop an example exercise
Implementation of scale forms
10.12.2018
Quizizz
Subject
Computer Networks
Applying the Quizizz test
Introduce the Quizizz app
Develop an example exercise
Implementation of scale forms
17.12.2018
Plickers
Subject
Information Technology Ethics
Applying the Plickers test
Introduce the Plickers app
Develop an example exercise
Implementation of scale forms

Figure 5. Implementation process

As it is shown in Figure 5, firstly the course subjects were explained. Then, the tests were applied in different types as paper-based, Kahoot, Quizizz, and Plickers app. In the next step, the application was introduced and students developed a practice at these applications. In the last stage, scale forms were applied to get the students' opinions about these practices. Detailed information about the findings is given in the findings section.

3. FINDINGS

This section contains detailed information about data analyses and findings. Firstly, descriptive information about the students who participated in the research are presented. Then, the findings obtained in the form of sub-problems is presented. Within the scope of this research, a four-week implementation was conducted with the students. Information about students who participated in each practice is presented in Table 2.

Table 2. Descriptive information about the students who participated in each practice

Division	Paper-based test	Kahoot	Quizizz	Plickers
Preschool Education	26	25	19	20
Guidance and Counseling	6	12	11	8
Social Sciences Education	29	23	29	26
Total	61	60	59	54

As it is shown in Table 2, 61 students participated in the paper-based test, 60 students participated in Kahoot, 59 students participated in Quizizz and 54 students participated in Plickers practices. The analyses were executed based on these students' responses. The findings are presented as sub-problems.

3.1. Findings about Perceived Usefulness

The first one of the hypotheses included in this study is to examine the perceived usefulness of the applications used. In order to examine this, one-way analysis of variance was performed. Before the analysis, normal distribution and homogeneity of variances were tested. The results of these analyses are presented in Table 3.

Table 3. Results of the normal distribution and homogeneity of variances perceived usefulness

Descriptive statistics for normal distribution		Homogenity of variance	
Skewness	-0,88	Levene statistic	5,28
Skewness std. error	0,14	Sig.	0,00*
Skewness/std error	-6,29		
Kurtosis	1,41		
Kurtosis std. error	0,28		
Kurtosis/std error	5,03		

*(p<0,05)

According to Table 3, it is seen that the assumptions required for performing one-way ANOVA (normal distribution and homogeneity of variance) are not ensured. In order to ensure these assumptions, Levene test results should be statistically insignificant and skewness/std. error and kurtosis/std. error values should be between +1,96 and -1,96 (Field, 2005). Transformations must be executed in order to provide the assumptions (Tabachnick, Fidell & Ullman, 2007). As a result of the analyses, it was seen that the assumptions were not provided and the square transformation was performed. Results about normal distribution and homogeneity of variances after transformation are presented in Table 4.

Table 4. Results of the normal distribution and homogeneity of variances after transformation for perceived usefulness

Descriptive statistics for normal distribution		Homogenity of variance	
Skewness	-0,08	Levene statistic	2,28
Skewness std. error	0,14	Sig.	0,60
Skewness/std error	-0,57		
Kurtosis	0,27		
Kurtosis std. error	0,28		
Kurtosis/std error	0,95		

p<0,05

As it is shown in Table 4, after the transformations, both normal distribution and homogeneity of variances were obtained. After this stage, one-way ANOVA was performed. Descriptive statistics were presented based on raw data. Square values were utilized for performing one-way ANOVA. The descriptive information of the analysis is presented in Table 5.

Table 5. The descriptive information about perceived usefulness

Application	N	$\overline{X}$	SD
Paper-based test	61	10,62	2,65
Kahoot	60	12,20	1,53
Quizizz	59	12,00	2,01
Plickers	54	11,11	1,90

As it is seen in Table 5, it can be said that the averages of different applications (paper-based test $\bar{X}$ =119,77; Kahoot $\bar{X}$ = 151,77; Quizizz $\bar{X}$ = 148,00; Plickers $\bar{X}$ = 127,03) are heuristically different. But, in order to use this expression, this situation needs to be statistically tested. The results of variance analysis are presented in Table 6.

Table 6. The results of variance analysis about perceived usefulness

	Sum of squares	df	Mean Square	F	Sig.	Differences
Between Groups	42614,34	3	14204,78	7,17	0,00	Kahoot>Paper-based test; Kahoot>Plickers Quizizz>Paper-based test
Within Groups	455885,05	230	1982,11			
Total	498499,39	233				

(*p*<0,05)

According to Table 6, it is seen that there is a statistically significant difference in perceived usefulness (F(3, 230)=7,17; p<0,05) according to the applications used in the classroom. In order to determine the difference between the groups, the Tukey test was conducted as a Post Hoc test. According to the results of the Tukey test, there is statistically significance. It was observed that a) the students' perceived usefulness to use Kahoot ($\bar{X}$ = 12,20; s = 1,53) instead of paper-based test ($\bar{X}$ = 10,62; s=2,65); b) the students' perceived usefulness to use Kahoot ($\bar{X}$ = 12,20; s = 1,53) instead of Plickers ($\bar{X}$ = 11,11; s = 1,90); and c) the students' perceived usefulness to use Quizizz ($\bar{X}$ = 12,00; s = 2,01) instead of paper-based test ($\bar{X}$ = 10,62; s=2,65) were statistically higher.

3.2. Findings about Behavioral Intention

The other hypotheses included in this study is to examine the behavioral intention of the applications used. In order to examine this, one-way ANOVA was performed. Before the analysis, normal distribution and homogeneity of variances were tested. The results of these analyses are presented in Table 7.

Table 7. Results of the normal distribution and homogeneity of variances behavioral intention

Descriptive statistics for normal distribution		Homogenity of variance	
Skewness	-0,73	Levene statistic	3,34
Skewness std. error	0,14	Sig.	0,01*
Skewness/std error	-5,21		
Kurtosis	0,50		
Kurtosis std. error	0,28		
Kurtosis/std error	1,79		

**p*<0,05

According to Table 7, as a result of the analyses, it was seen that the assumptions were not ensured and square transformation was performed. Results about normal distribution and homogeneity of variances after transformation is presented in Table 8.

As it is shown in Table 8, after the transrormations, both normal distribution and homogeneity of variances were obtained. After this stage, one-way ANOVA was performed. Descriptive statistics were presented based on raw data. Square values were utilized for performing one-way ANOVA. The descriptive information of the analysis is presented in Table 9.

Table 8. Results of the normal distribution and homogeneity of variances after transformation for behavioral intention

Descriptive statistics for normal distribution		Homogenity of variance	
Skewness	0,06	Levene statistic	1,95
Skewness std. error	0,14	Sig.	0,10
Skewness/std error	0,43		
Kurtosis	-0,29		
Kurtosis std. error	0,28		
Kurtosis/std error	-1,04		

p<0,05

Table 9. The descriptive information about behavioral intention

Application	N	$\overline{X}$	SD
Paper-based test	61	9,86	3,37
Kahoot	60	11,55	2,21
Quizizz	59	11,28	2,43
Plickers	54	10,50	2,60

As can be seen in Table 9, it can be said that the averages of different applications (paper-based test $\overline{X}$ =108,59; Kahoot $\overline{X}$ = 138,22; Quizizz $\overline{X}$ = 133,25; Plickers $\overline{X}$ = 116,91) are heuristically different. But, in order to use this expression, this situation needs to be statistically tested. The results of variance analysis are presented in Table 10.

Table 10. The results of variance analysis about behavioral intention

	Sum of squares	df	Mean Square	F	Sig.	Differences
Between Groups	34356,66	3	11452,22	3,88	0,1	None
Within Groups	678522,66	230	2950,01			
Total	712879,32	233				

(p<0,05)

According to Table 10, it is seen that there is not a statistically significant difference in behavioral intention (F(3, 230)=3,88; p>0,05) according to the applications used in classroom.

4. DISCUSSION and CONCLUSION

CRSs provide feedback to instructors on what subjects are needed for extra teaching and in which topics students are successful (Bartsch & Murphy, 2011). Besides, CRSs give learners immediate feedback on multiple-choice questions (Judson & Sawada, 2002). This feedbacks provide important tips for formative assessment. In order to use these systems, in the first step teachers and candidate teachers should be informed about these systems or environments. In the second step, they should think that these systems are useful and they want to use them. So, in the context of this research, the students had an experience with different CRSs as Kahoot, Quizizz, and Plickers. And then students' perceived usefulness and behavioral intention structures about the CRSs were investigated. In the literature, it is concluded that the use of CRS is easy to use, useful by learners, increasing interaction and engagement (Siau et al., 2006; Martyn, 2007; King & Robinson, 2009; Sievers et al., 2012; Keyssner & Dunn, 2013; Wu, Wu, & Li, 2019). CRS tools provide immediate feedback (Kay, 2009). Black and William (1998) stated that quality feedback is the key feature of FA. So, these tools can be utilized as a FA tool in the classroom. And also, more enjoyable and competitive classroom climate can be created through CRS.

This research was conducted in a computer laboratory with internet connection. So, there was no problem with the internet connection or physical infrastructure. CRSs are technologies that are used to promote active learning (Martyn, 2007). So, learning environments should be structured to encourage students and increase students' engagement. And also, learners must be an active part of the learning process in these environments. However, it should be noted that the essential hardware and physical infrastructure should be provided when using these systems because the inability to provide productive classroom environments is one of the problems encountered in the effective use of these systems (Feldman & Capobianco, 2003). If the researcher/instructor/tutor wants to use CRS, it is important that each student must have a mobile device or computer with an internet connection. In addition, if the questions will be displayed to the students on a single screen, the screen should be positioned where all the students can see the questions.

Nowadays the CRSs present much detailed information to the instructors such as a) individual performance, b) performance of the group, c) correct response rate on for each question, and d) the response time of each question. This information is very important for the application. Based on this information, environments or applications can be reconstructed or reorganized. And for these systems, one of the usage points to be considered is to give the student enough response time (Lucke, Keyssner, & Dunn, 2013). Otherwise, the effectiveness of the learning environment will be affected negatively. Besides, the addition of the elements of gamification as a leader board, and badges provides opportunities for the learning environment to make it more enjoyable and to see their status according to their peers. In addition to these, the studies and applications should be integrated well into the curriculum (Sims, 2003).

Within the scope of this study, perceived usefulness and behavioral intention of using CRS were examined by comparing both the paper-based test and each other unlike studies in the literature. This research is limited to perceived usefulness and behavioral intention structures of the acceptance and adoption theories and models. For future research, the other structures of the acceptance and adoption theories and models as perceived ease of use, social impact, content and etc. structures could be examined. In addition to this, these structures can be tested by structural equation modeling.

This study was structured as survey research. And findings are limited to the self-report scale data. In this context for future research, experimental design can be constructed or organized. In this way, the effectiveness of CRSs systems in the context of formative assessment can be demonstrated more clearly. The other suggestion is to conduct this research with a higher number of students/participants group. Thus, it may make possible to present more clear findings which is generalizable to the population.

ORCID

Muhittin Şahin https://orcid.org/0000-0002-9462-1953

5. REFERENCES

Bartsch, R. A., & Murphy, W. (2011). Examining the effects of an electronic classroom response system on student engagement and performance. *Journal of Educational Computing Research*, *44*(1), 25-33.

Bayrak, F., & Yurdug ül, H. (2016). Web-tabanlı öz-değerlendirme sisteminde öğrenenlerin öz-müdahale algisi ve test alma davranişlarinin başari üzerine etkisi [The effect of self-intervention perception and test taking behaviour on success in web-based self-assessment system]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *7*(1), 221-236.

Bayrak, F., & Yurdugül, H. (2016). Web-tabanlı öz-değerlendirme sisteminde öğrenci uyarı indeksini temel alan öğrenme analitiği modülünün tasarlanması [Designing at the micro level learning analytics module for web-based self-assessment system]. *Eğitim Teknolojisi Kuram ve Uygulama, 6*(2), 85-99.

Beatty, I. D., Gerace, W. J., Leonard, W. J., & Dufresne, R. J. (2006). Designing effective questions for classroom response system teaching. *American Journal of Physics, 74*(1), 31-39.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). "How People Learn: Brain, Mind, Experience, and School", Expanded Edition, Washington, DC: National Academy Press.

Cheng, L. T., & Wang, J. W. (2019). Enhancing learning performance through classroom response systems: The effect of knowledge type and social presence. *The International Journal of Management Education, 17*(1), 103-118.

Deal, A. (2007). A teaching with technology white paper: classroom response systems. Office of Technology for Education. Eberly Center for TeachingExcellence. Retrieved June, 25, 2009.

Feldman, A., & Capobianco, B. (2003). Real-time formative assessment: A study of teachers' use of an electronic response system to facilitate serious discussion about physics concepts. In Annual Meeting of the American Educational Research Association (Chicago, IL, 2003).

Field, A. (2005). *Discovering statistics using SPSS*. Sage Publication Ltd.

Fishbein, M., & Ajzen, I. (1975). Belief, attitude, intention and behavior: an introduction to theory and research: Addison-Wesley, Reading MA.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2011). How to design and evaluate research in education. New York: McGraw-Hill Humanities/Social Sciences/Languages.

Fuller, J. S., & Dawson, K. M. (2017). Student response systems for formative assessment: Literature-based strategies and findings from a middle school implementation. *Contemporary Educational Technology, 8*(4), 370-389.

Judson, E., & Sawada, D. (2002). Learning from past and present: Electronic response systems in college lecture halls. *Journal of Computers in Mathematics and Science Teaching, 21*(2), 167-181.

Kay, R. H. (2009). Examining gender differences in attitudes toward interactive classroom communications systems (ICCS). *Computers & Education, 52*(4), 730-740.

Kay, R. H., & LeSage, A. (2009). Examining the benefits and challenges of using audience response systems: A review of the literature. *Computers & Education, 53*(3), 819-827.

King, S. O., & Robinson, C. L. (2009). 'Pretty Lights' and Maths! Increasing student engagement and enhancing learning through the use of electronic voting systems. *Computers & Education, 53*(1), 189-199.

Lucke, T., Keyssner, U., & Dunn, P. (2013, October). The use of a classroom response system to more effectively flip the classroom. In 2013 IEEE Frontiers in Education Conference (FIE) (pp. 491-495). IEEE.

Martyn, M. (2007). Clickers in the classroom: An active learning approach. *Educause Quarterly, 30*(2), 71.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory. New York: McGraw-Hill.

Petto, A. J. (2019). Technology meets pedagogy: Comparing classroom response systems. *Journal of College Science Teaching, 48*(4), 55-63.

Pinchok, N., & Brandt, W. C. (2009). Connecting formative assessment research to practice: an introductory guide for educators. Learning Point Associates.

Rogers, E. (2003). *Diffusion of innovation*. New York: Free Press.

Siau, K., Sheng, H., & Nah, F. H. (2006). Use of a classroom response system to enhance classroom interactivity. *IEEE Transactions on Education*, *49*(3), 398-403.

Sievers, M., Reinhardt, W., Kundisch, D., & Herrmann, P. (2012). Developing electronic classroom response apps for a wide variety of mobile devices: Lessons learned from the PINGO project. In mLearn (Vol. 955, pp. 248-251).

Sims, R. (2003). Promises of interactivity: Aligning learner perceptions and expectations with strategies for flexible and online learning. *Distance Education*, *24*(1), 87-103.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). Using multivariate statistics (Vol. 5). Boston, MA: Pearson.

Usluel, Y. K., & Mazman, S. G. (2010). Eğitimde yeniliklerin yayılımı, kabulü ve benimsenmesi sürecinde yer alan öğeler: bir içerik analizi çalışması. *Çukurova University Faculty of Education Journal*, *3*(39), 60-74.

Vetterick, J., Garbe, M., Dähn, A., & Cap, C. H. (2014). Classroom response systems in the wild: technical and non-technical observations. *International Journal of Interactive Mobile Technologies (iJIM)*, *8*(1), 21-25.

Wu, Y. C. J., Wu, T., & Li, Y. (2019). Impact of using classroom response systems on students' entrepreneurship learning experience. *Computers in Human Behavior*, *92*(2019), 634-645.

Yang, X., Fu, J., Rodríguez-Sedano, F., & Conde, M. Á. (2019, October). Enhancing Students' Academic Performance through Teamwork and Classroom Response Systems: Case Study with Chinese Students. *In Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 732-737). ACM.

Yılmaz, R., & Karaoğlan Yılmaz, F. G. (2019). Bir oyunlaştırma ve biçimlendirici değerlendirme aracı olarak kahoot kullanımına yönelik öğretmen adaylarının görüşlerinin incelenmesi. *II. Uluslararası Eğitimde ve Kültürde Akademik Çalışmalar Sempozyumu (331-337).* Denizli, Türkiye

Yurdugül, H. & Bayrak, F. (2014). İlkokul öğrencilerinin web tabanlı biçimlendirmeye dönük değerlendirme sistemini kabulleri [The acceptance of web based formative assessment system for primary school students]. *Journal of Educational Sciences & Practices, 2014,* 13-26.

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 706–720
https://dx.doi.org/10.21449/ijate.548516

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

Developing a Scale to Measure Students’ Attitudes toward Science

Adem Akkuş [1,*]

[1] Mus Alparslan University, Education Faculty, Elementary Science Education, Mus, Turkey

ARTICLE HISTORY

Received: 03 April 2019
Revised: 31 October 2019
Accepted: 06 December 2019

KEYWORDS

Science Attitude,
Scale Development,
Scale validation

Abstract: The aim of this study is to develop a science attitude scale (SAS). For that purpose, the literature review has been done for suggestions for creating scales and a new draft scale developed. The draft scale was analyzed by specialists and a pilot study is done after its approval by experts. The SAS is prepared with 21 items and among these, 11 items are reverse-coded. The SAS consists of Likert-type items. The sample of the study consists of 154 college students studying at the Faculty of Education, Elementary Science Education, and Elementary Education departments. Principal axis factoring with orthogonal rotation (varimax) was used for exploratory factor analysis. Factor eigenvalues were checked with respect to parallel analysis and numbers of the factors were determined with respect to the analysis. Items that did not serve the purpose of the scale were omitted from the SAS. The finalized SAS’ Cronbach alpha value is .953. For confirmatory factor analysis data were collected from a different sample which consists of university students who were studying at elementary science education, elementary education, and electric electronic engineering departments. Number of sample is 201. Confirmatory factor analyses run through Amos 24.0 software. It is believed that SAS is a valuable contribution to the science education field since it has unidimensional structure and proved its item discrimination power, and alongside with an excellent internal consistency. SAS also offers opportunity to develop multidimensional science attitude scale. For that purpose, original SAS and English version of it are provided in appendixes.

1. INTRODUCTION

Attitude is defined as an individual’s positive or negative characteristics towards a subject (Serin & Mohammadzadeh, 2008). Students with positive attitudes toward science are likely to display more science-related attitudes and choose science-related professions. On the other hand, recent studies indicate that there is a trend in that science-related departments attract fewer students than social science-related departments (Shah & Mahmood, 2011). Therefore, attitudes toward science and science-related subject areas are in focus of research studies. Even science attitudes may be used for predicting science achievement (Adesoji, 2008). Factors affecting attitudes are also among the subjects to be studied. For example, gender might be suggested as one of the factors. Although both genders have closely similar attitude values

CONTACT: Adem AKKUŞ ✉ ademakkus@gmail.com Mus Alparslan University, Education Faculty, Elementary Science Education, Mus, TURKEY

ISSN-e: 2148-7456 /© IJATE 2019

toward science, underlying factors might be different. Girls learn better in an organized environment and boys' attitudes are related to cohesiveness. Other factors might be listed as instructional style, teaching strategies, classroom design, etc. (Bernardez, 1982). Thus, having knowledge of students' attitudes and encouraging them toward science is important and attitudes of students must be known (Shah & Mahmood, 2011). Scales are useful for this purpose and in this regard, researchers try to develop their own instruments for various purposes or use a standardized version (Coll, Dalgety & Salter, 2002). Using standardized scales or other means of standardized measurements could value the purpose, letting researchers have an opinion on the attitudes of students and understand dimensions and their value within the context (Demirbaş, 2009). On the other hand, standardized scales are mostly in English and have different theoretical aspects, different cultural settings, psychometric properties and hence may lack assess the right domain of interest, not be suitable for local use due to contextual differences (Shah & Mahmood, 2011). Perhaps that is the reason why different researchers have failed to confirm the Test of Science Related Attitudes scale (TOSRA) in their sample population. Attitudes may be observed in different types of responses and even be affected by curriculum changes (Cheung, 2007). As curriculum changes are made, the need for to measure attitudes and to develop new scales also becomes a value of interest to observe the effect of the curriculum. Even instructional techniques might affect students' attitudes whose change could value the future implications (Evrekli, İnel, Balım & Kesercioğlu, 2009). Since Turkey has already announced that changes on the curriculum are done to promote active learning (TTK, 2017), it is important to observe the effects of curriculum changes on students' attitudes. For that aim, several researchers already tried to develop attitude scales or applied existing ones. For example, Can and Şahin (2015) studied kindergarten teacher candidates' attitudes toward science and science teaching. Analyses were done to investigate the relationships of grades and gender with science attitude and science teaching. Serin & Mohammadzadeh (2008) used a scale to determine attitude and academic success relationships. Korkmaz, Şahin and Yeşil (2011) tried to investigate attitude toward scientific research. For that reason, they developed a scale with 30 items and four dimensions. Tortop (2013) adapted a scale into Turkish for assessing scientific field trip attitude. The study reveals that a single attitude might have different dimensions. For example, another study tried to investigate the relationships between attitudes and science process skills (Dönmez & Azizoğlu, 2010). All these studies show that scales might be used for collecting data (Deshpande, 2004; Hinkin, 1998; Wong & Lian, 2003; Francis et. al., 2004) so that effective measures might be taken into account for this purpose (Hinkin, 1998; Hinkin, Tracey, Enz, 1997). Thus, the purpose of this study is to develop a science attitude scale (SAS). Attitudes may have different dimensions and scales may reflect those dimensions. However, most scales determined the number of dimensions based on eigenvalues through factor analysis. SAS also determined number of dimensions through parallel analysis which reflects more accurate number of dimensions. Moreover, SAS is a unidimensional scale but offers the chance for researchers to develop multidimensional scales based on SAS.

2. METHOD

2.1. Research Design

In the method, to develop a scale, based on suggestions from the literature, the guidelines have been determined (Brinkman, 2009; Deshpande, 2004; Hinkin, 1998; Hinkin et. al., 1997; Johanson & Brooks, 2010; Ajzen, 2005b; Francis et. al., 2004; Cabrera-Nguyen, 2010; Hof, 2012). Those guidelines are:

a) Not to cause any bias, the items' context must be within the students' cultures (schemes).
b) Respondents should place themselves at a position.

c) Items must assess a single behavior or response.
d) Items must not be interpreted in different ways.
e) Language and expected knowledge should be familiar with the target group.
f) Sensitive and double negative items should be avoided.

Among Thurstone's method of equal-appearing intervals, Likert scale, semantic differential scales; it is determined that a Likert type scale would be more beneficial for the purpose of the study to ensure easy compilation and generalization (Lovelace & Brickman, 2013; Brinkman, 2009; Johanson & Brooks, 2010; Hof, 2012). To ensure content adequacy and avoid fatigue, a maximum number of items is determined so that respondents will respond within attention time. For this purpose, the sentences "Strongly agree" or "Strongly disagree" are given at the beginning of the scale as information. By placing five levels of response for an item, it is ensured that internal consistency is increased and sufficient variances are obtained (Hinkin, 1998; Lovelace & Brickman, 2013; Brinkman, 2009; Hinkin et al., 1997; Ajzen, 2005b; Francis et al., 2004). Since there might be respondents tending to choose options at the edges or in the same direction, reversed coded sentences are appropriately used to trigger their vigilance (Hinkin, 1998; Francis et al., 2004; Hof, 2012)

The process of developing the science attitude scale (SAS): The item sentences were finalized after determining SAS' scope, content, items and their numbers. After that specialist views were taken account. The draft science attitude scale (SAS) consisted of 31 items. However, items 25 and 27 were removed from the SAS since they were the same as items 6 and 5. Initial internal reliability analysis was carried out and Cronbach's α value was found as .861 (good according to Kalaycı, 2010). The draft scale's content and scope were analyzed by instructors who have the experience of teaching and researches on related issues since specialists could value the prepared scale on the content domain. Specialists work in the education faculty at science education department (Hinkin et al., 1997). Specialists' views' on sentences and corrected item-total correlation values of the items were cross-checked, and the items regarded as problematic were excluded from the scale. Item of 1 and 9 contradicted guideline f "Sensitive and double negative items should be avoided", items of 8, 11, 14, and 26 contradicted guideline c "Items must assess a single behavior or response" and guideline d "Items must not be interpreted in different ways". Thus, those items were excluded from the SAS immediately for further analysis; thus finalized SAS Cronbach's α reliability value is .943 and with 23 items. The developed SAS consists of twelve reversed questions (items) which are items 2, 5, 6, 7, 10, 16, 17, 20, 21, 22, 23, and 31.

2.2. The sample size and sampling method

The SAS was applied to 154 college students at the Faculty of Education, Elementary Science Education, and Elementary Education departments. In order to ensure the privacy of personal information, (i.e. avoiding conflict of interest) only the students' gender and age information were demanded.

In literature, to determine a sample size has been a controversial issue. Some researchers argue about arbitrary sampling which presents high communalities without cross-loadings. Thus, sampling may be determined by the nature of the data i.e. More acceptable view, some researchers claim that if data is strong enough then sample size might be small, while others argue on item-ratio sampling. The debate on item-ratio suggests proportion from 1:2 to 1:10 for item and sampling (Anthoine, Moret, Regnault, Sébille & Hardouin, 2014; Hinkin, 1998; Hinkin et al., 1997; Cabrera-Nguyen, 2010). Since the item respondents' ratio of the study is 1:7, it is believed that sampling is adequate for the study with respect to first view.

As for arbitrary sampling, several arguments might be stated. For example, Johanson & Brooks (2010) suggests to social researchers that minimum participants are 100 people for sampling.

For a comprehensive item analysis, sample size 100 to 200 person should be conducted since standard errors for Cronbach's alpha value increase as the sample size decreases. However, it is also noted that regardless of the number of items, the mean inter-item correlation is nominal between N= 30 and 200. Hinkin (1998) and Hinkin et al., (1997) suggest N=150 to obtain sufficient data for exploratory factor analysis as long as item inter correlations are reasonably strong. However, researchers also note the difference between statistical and practical significance because attaining statistical difference chance increases as the sample size increases. Larger samples are useful to detect small fluctuations. On the other hand, as sample size increases the practical meaning of the results may distort, so the decision on sample size must be made with caution. Francis et al. (2004) state that some researchers claim that N=25 would be enough for purposive sampling and sample size could be increased until it is believed that data saturation has been achieved. Yet, researchers claim a sample size N=80 would generally be enough. Cabrera-Nguyen (2010), while stating similar statements, also indicates that some researchers argue on sampling size and claim sample size depends on the gathered data, and adequacy of sampling is determined after analyzing the gathered data. In the same paper, the researcher embraces a mixed approach based on communalities value. Hof (2012) suggests 10-15 respondents per item, yet states a KMO value already signals whether the sample size is enough or not. Based on the suggestions above, it is again regarded that the sample of the study will be enough for the purpose.

3. RESULT

3.1. Reliability Analysis

The data analyzed with respect to internal consistency, communalities, and factor loadings. Analyses were carried out together for better judgment of retaining factors. Corrected item-total correlation values of the items were analyzed. It was observed that items 12 (.004) and 16's (.126) corrected item-total correlation values were below the desired value of .200 (Johnson & Morgan, 2016), hence; these items were excluded from the scale. After this process the scale's Cronbach's α value was found as .953 and "excellent" for the final version of SAS.

3.2. Exploratory Analysis

Exploratory Factor Analysis: A principal axis was conducted on the 21 items with orthogonal rotation (varimax) through the SPSS program to reveal the factors within the developed scale since it is suggested for more reliable scale evaluation (Field, 2013; Hof, 2012). The Kaiser-Meyer-Olkin (KMO) measure verified the sampling adequacy as "marvelous" (Kalaycı, 2010). The KMO value is .951 and above the acceptable limit of .5 (Field, 2013). Bartlett's test of sphericity was found significant ($X^2(210) = 2392.067$, p= .00 < .05). Hence, the KMO value already signaled that the sample size might be enough, the analysis of each SAS item was initiated. An initial analysis was run to obtain eigenvalues for each factor in the data. Three factors emerged having eigenvalue over Kaiser's criterion of 1 and in combination explained %58,778 of the variance. Eigenvalue of the factors were 11,301; 1,340 and 1,017 respectively for factor 1, factor 2 and factor 3.

The parallel analysis offers a good interpretation of the number of retaining factors (Field, 2013; Johnson & Morgan, 2016) thus; a Monte Carlo PCA for parallel analysis with 1000 replications was run to confirm the eigenvalues (Watkins, 2000). Eigenvalues obtained were 1,7259; 1,5915 and 1,4902 for that reason, it was concluded that only the first eigenvalue was acceptable since the second and third factor's eigenvalue was not significant. Therefore, an EFA was rerun with a one-factor solution. The variance shared by the factor was 51,937. The scree plot (Figure 1) was obtained and it was decided that the scale has one factor with respect to the convergence of the scree plot and Kaiser's criterion on this value. Table 1 shows the factor loadings after rotation and extracted communalities.

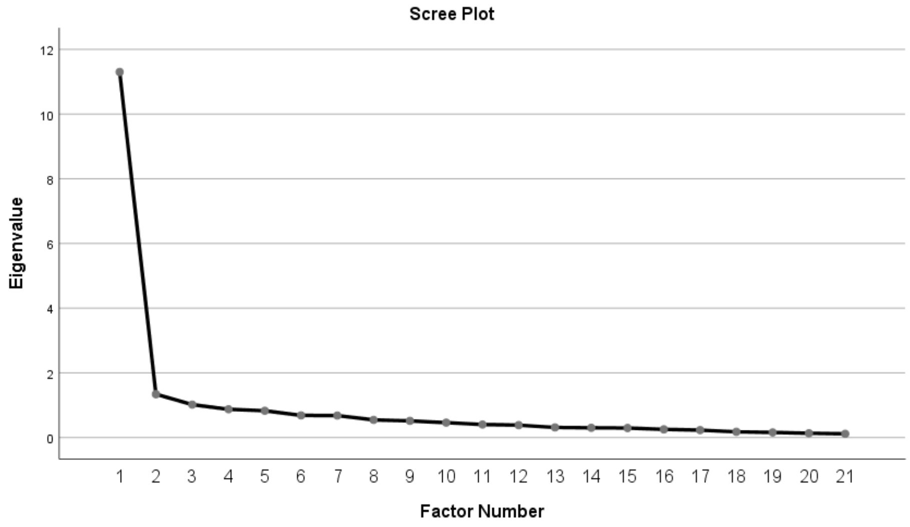


Figure 1. Scree Plot

Table 1. extracted communalities and factor loadings

Item	h^2	Factor
SAS2	,215	,464
SAS3	,602	,776
SAS4	,568	,754
SAS5	,422	,650
SAS6	,474	,689
SAS7	,259	,509
SAS10	,174	,417
SAS13	,494	,703
SAS15	,787	,887
SAS17	,621	,788
SAS18	,295	,543
SAS19	,826	,909
SAS20	,773	,879
SAS21	,549	,741
SAS22	,671	,819
SAS23	,721	,849
SAS24	,624	,790
SAS28	,207	,455
SAS29	,817	,904
SAS30	,617	,785
SAS31	,190	,435

Since none of the items' factor loadings were below .400 (Table 1) validation of scale's internal consistency reliability coefficient was made as suggested (Field, 2013; Francis et. al., 2004). Cronbach's α value was found as =.953 and it means "excellent" (Kalaycı, 2010). For a detailed analysis of items on discrimination value, an independent samples t-test was run for each item. Lower and upper %27 of the samples (N=42) were compared through independent samples t-test. This analysis shows items' discrimination value of individuals between lower and upper %27 of the sample and is used by many scientists in scale developments (Moore & Foy, 1997). Reliability analysis of item-total correlation and $t_{up-down(\%27)}$ results and items' codes are shown in Table 2.

Table 2. Item-total correlation and $t_{up-down(\%27)}$ results

Item	Mean	Standard Deviation	Corrected-item total correlation	$t_{up-down(\%27)}$
SAS1	2,92	1,603	,456	5,767*
SAS2	3,14	1,500	,757	12,781*
SAS3	3,05	1,522	,738	12,796*
SAS4	3,14	1,650	,643	9,729*
SAS5	3,18	1,602	,677	14,399*
SAS6	3,23	1,475	,506	6,588*
SAS7	3,06	1,423	,412	4,944*
SAS8	3,07	1,417	,680	9,175*
SAS9	3,31	1,619	,863	26,213*
SAS10	3,20	1,479	,765	12,029*
SAS11	3,06	1,538	,520	7,381*
SAS12	3,19	1,555	,889	22,715*
SAS13	3,38	1,500	,862	23,059*
SAS14	3,03	1,434	,722	13,191*
SAS15	3,17	1,564	,801	16,176*
SAS16	3,26	1,385	,829	17,507*
SAS17	3,22	1,515	,771	13,456*
SAS18	3,16	1,380	,444	5,343*
SAS19	3,19	1,443	,887	20,686*
SAS20	3,08	1,412	,764	14,565*
SAS21	3,01	1,186	,414	6,093*

* $p < .05$

The finalized SAS consists of 21 items and 11 items are reverse coded items. The reversed coded items are 1, 4, 5, 6, 7, 10, 13, 14, 15, 16 and 21 (Table A1). For international readers, an English translation of SAS is given in Table A2. Translation was done by the researcher and to ensure translation was done correctly and comprehension of the scale is easy, SAS was presented to a professor to take account of specialist's opinion. That professor was working at university and had a formal education in English language. After that a retranslation and crosscheck were done by another professor who also had a formal education in English language and working at Education faculty.

3.3. Confirmatory Factor Analysis

For confirmatory factor analysis data were collected from a different sample. The sample consists of university students who were studying at elementary science education, elementary education, and electric electronic engineering departments. The number of sample is 201 in total. Confirmatory factor analyses run through Amos 24.0 software. Initial analysis results revealed that $\chi 2$/DF ratio is 2,635 RMSEA value is .09; GFI value is .802; CFI value is .627; SRMR value is .0852; NFI value is .520; AGFI value is .758. RMSEA, GFI, AGFI values showed model did not show a good fit with respect to indices values. However, it is noteworthy that they are close to the desired value. CFI and NFI values did not fall into the categories of well fit. Thus, in the review of the literature it is decided to examine the error terms and decide whether some items correlate together or not for a better comprehension of the model. Decisions to covariate items based on the rule of thumbs. First, covariated items should exhibit similar set of idea/pattern/mindset in phrases. Second, number of covariated items should be restricted up to 7. Third, each time a covariate is done then, the results would be checked to see if model fit indices were changed dramatically or not. Fourth, covariances between the items should not be lower than 10. Covariated items are 13 and 14; 9 and 11; 12 and 17; 18 and 19. Final data analyses results revealed that χ2/DF ratio is 1.893; RMSEA value is .067; SRMR value is .0746; IFI value is 0.806; GFI value is .852; CFI value is .800; NFI value is .662; PNFI value is 0.583; NNFI (TLI) value is 0.773; AGFI value is .815. The confirmatory factor analysis result is shown in Figure 2.

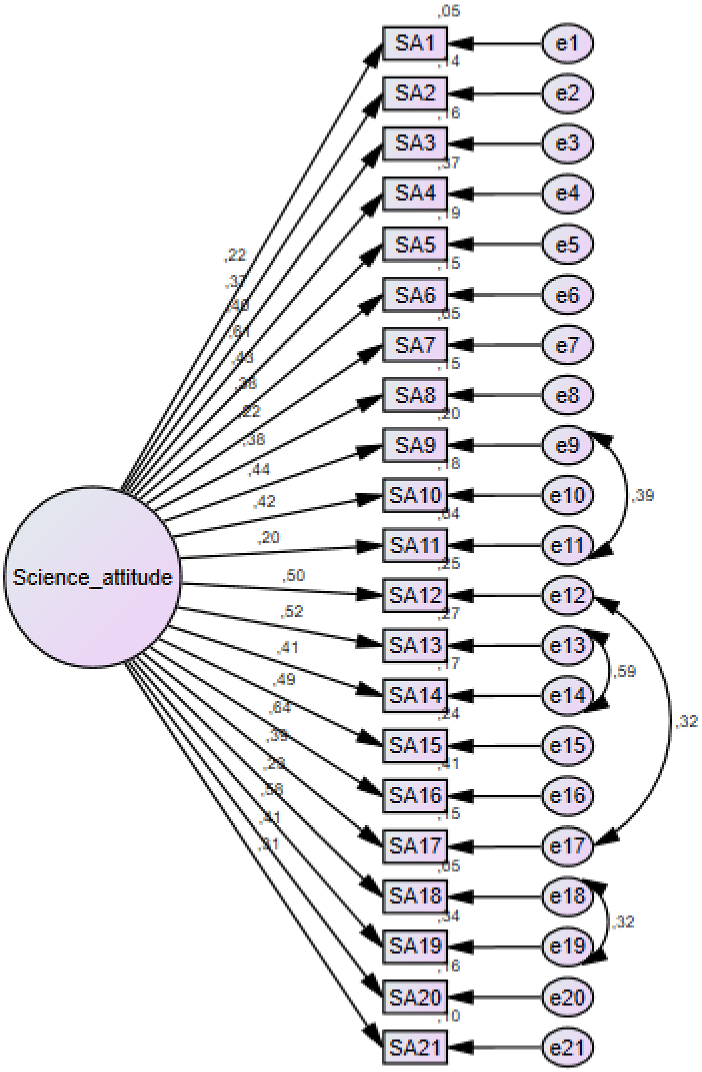


Figure 2. Confirmatory Analysis Result

χ2/DF ratio is 1.893 and it is regarded that a model has a good fit if Chi-square (χ2)/degree of freedom (df) ratio is < 2. RMSEA value is .067 and model has a good fit since RMSEA ≤ 0.1. IFI value is 0.806 and CFI value is .800 and it is accepted model has a good fit since CFI ≥ 0.8 and IFI ≥ 0.8. (Browne and Cudeck, 1993; Garson, 2006 as cited in Chinda, Techapreechawong & Teeraprasert, 2012).

Many recommendations are being done on value of Root Mean Square Error of Approximation (RMSEA). For example, Pedroso et. al. (2016) state if RMSEA is ≤ 0.05 then, it indicates good fit and if RMSEA is ≤ 0.08 it indicates good fit with reasonable errors. Other recommendations generally advise that if RMSEA ≤ 0.08 then a model has good fit and, if 0.08 < RMSEA ≤ 0.1

then it indicates adequate fit (Carlback & Wong, 2018; Shadfar & Malekmohammadi, 2013). The confirmatory analysis revealed that RMSEA value is .067 so SAS has good fit of model.

Since RMR has bias, SRMR used instead of RMR. SRMR value of the confirmatory factor analysis result is 0.0746. SRMR value below ≤ 0.08 indicates good fit (Carlback & Wong, 2018; Kline, 2011 as cited in; Kaya & Altinkurt, 2018; Vassallo & Saba, 2015).

Kline (2011) mentions that most used fit indices (GFI, AGFI, NFI, NNFI, CFI and IFI) should be ≥ 0.85 (cited in; Kaya & Altinkurt, 2018; Vassallo & Saba, 2015). Other researchers indicate AGFI, GFI and CFI values should be ≥ 0.80 (Byrne & Campbell, 1999 as cited in Nayir, 2013). However, GFI is affected by sample size and for that reason, AGFI is developed. GFI ≥ 0.85 and AGFI ≥ 0.80 is accepted as good fit (Sica & Ghisi, 2007). On the other hand, AGFI also is sensitive to sample size. For that reason, it is advised by researchers to disregard them. Yet papers still indicate both values. Reason for that is not for their importance but historical values. TLI ≥ 0.85 indicates good fit and > 0.8 mediocre fit (Carlback & Wong, 2018; Shadfar & Malekmohammadi, 2013). The confirmatory analysis revealed that IFI value is 0.806; GFI value is .852; CFI value is .800; NNFI (TLI) value is 0.773; AGFI value is .815 so SAS has good fit of model.

4. DISCUSSION and CONCLUSION

SAS item discrimination values show that it might be used to measure the attitude toward science of college students since all the items yielded significant results between up-down%27. On the other hand, a detailed analysis of t-test values might reveal the facts lying beneath. For example, the highest t value (26,213) of item9 implies that students who have high positive attitudes enjoy the experiments, however; the value of items11 and 21 (7,381 and 6,093) imply students wait for confirmation on their experiment results or expect more guidance during experiments. Fin (2012) states perceived learning increases as students get feedback from instructors, thus; the expectation of students is meaningful and expected in this context. Similar reports indicate that instructor confirmation has a positive effect on cognitive learning (Schrodt, Witt, Turman, Myers, Barton & Jernberg, 2009). Therefore, instructors need to interact with students and help them in cognitive development.

Students have a tendency of thinking that feelings might assist a scientist. Low t value (4,944) of item 7 already implies that both upper-lower (%27) students think that facts may have subjective aspects. Perhaps, students think that scientific facts may change, and some unscientific factors act as catalysts for that change. In fact, t value (5,767) of SAS1 already hints that both upper-lower (%27) students have a tendency of the idea that facts might be subjective. From a different aspect, that approach might be seen through t value (6,588) of SAS6. That item indicates faith plays an important role in students' ideas about science's role. Conflicts between the dimensions may create gaps for the students on the nature of science. As a result, students struggle between scientific facts and faith. Since faith requires believing in what is being told without question, it may also create a barrier toward scientific approach and hence, students may close themselves to new ideas. Students who have high positive attitudes toward science are also open to new ideas and t value (22,715) of SAS12 already reveals this. One of the purposes of science courses is learning how to distinguish faith from science and instructors should make students aware that; two concepts actually are not related. In fact, faith and science have different roles and do not need to cut out each other's way. For instance, a high value of SAS19 (t =20,686) clarifies that students who have higher positive attitudes also acknowledge that scientific knowledge is essential as it is related to life itself. Embracing this idea might eventually increase positive attitudes toward science. Similarly, Nuhoglu (2008) who developed a science attitude scale in Turkey mentions that one of the factors is "new knowledge and using it" and Ajzen (2005a) points out that changing the attitudes and behaviors may be achieved through changing beliefs.

For a cross-cultural analysis, Scientific Attitude Inventory II (SAI-II) was also analyzed since some statements are similar to the statements in SAI-II. However, researchers did not provide an item t%27up-down values thus a comparison could not be made (Moore & Foy, 1997). Similarly, another well-known scientific attitude scale TOSRA, which has seven dimensions, also did not provide such an approach. On the other hand, the author states that having a low score on one of the dimensions should not concern instructors because this information might be valuable in facilitating to identify profiles and creating solutions. One of the most important aims of science education is considered to develop a positive attitude towards science (Fraser, 1981).

Motivation towards science has a long-term effect on science learning and it is affected by different things such as curriculum structures. For that reason, Foley & McPhee (2008) investigated the effect of different curriculum approaches on their study and argued that students' experiences might be affected by different curriculum structures. Kurnaz and Yigit (2010) report that Turkish students have tended to develop a negative attitude towards science since 2005. Thus, it is important to assess the changes of attitudes caused by curriculum changes in Turkey (TTK, 2017). Although this research was done on local scale, it might be said that science attitude affects the scientific approach and the scientific approach is the same all around the world. Moreover, attitudes might be affected by a similar insight context whose effect may yield similar results. Perhaps through such studies, identifying and creating solutions will be possible. Researchers and teachers might use the developed scale and observe their students', attitudes, use interventions and may offer solutions. Then, perhaps understanding students' attitudes on different dimensions may also offer solutions for long term aims.

An advantage of this study is seen as using parallel analysis to confirm eigenvalues in identifying dimensions. Since the traditional factor analysis determination procedure is based on eigenvalue of 1 then, the obtained number of factors may not be accurate. It is believed through such analysis such cases are avoided and true factor structures are determined in creating the SAS (Hayton, Allen & Scarpello, 2004). Developing science attitude scales is important, even restudying the existing ones may provide useful information. For that reason, researchers either create their own scales or develop the existing ones (Moore & Foy, 1997). Having an excellent internal consistency (α=.953) and a consisted structure (one factor structure) measuring the related domain, it is believed that SAS is a valuable contribution to the science education field.

Validity must be considered each time when an instrument is used since the instrument was validated for a sample or population but was not validated for another sample or population i.e. structure may show varying results from a sample to another. Validity is not property of a scale but it means as an instrument of interpretation. There are arguments on cut off values of fit indices since they may lead the decision of an acceptable model to be rejected. Thus, it is important for researchers to conduct the analyses and use their own judgments with respect to obtained values. Values of fit indices help to understand the structure of a model and thinking all the fit values together will provide a better decision. In other words, fit indices will help to understand relationships of the items among each other and within the model structure. Fit indices should confirm the model but it should not be used for championing the model in every possible indices which will cause an artificially approved model. It is advised that once items and factors make sense in the theoretical aspect of the researcher then, the decision will be based on that (Knekta, Runyon & Eddy, 2019). Hu & Bentler (1995) argue decisions based on fit indices and reminds that strictly depending on fit indices values may result in rejecting true models, especially for small sample sizes such as 250 or 500. GFI and AGFI tend to increase when sample size increase same thing could be also said for RMR and RHO (Anderson & Gerbing, 1984). NFI results could be problematic if sample size is < 200 thus usage of NNFI (TLI) is recommended. However, it is also noted that even NNFI could still indicate poor fit if sample size is not enough. Similar arguments are proposed for TLI since it could report poor fit due to similar reasons. For that reason, suggestions for TLI could be as low as 0.80 and, for parsimony fit indices (PNFI) values ≥ 0.5 indicates good fitness (Hooper, Coughlan & Mullen, 2008). The confirmatory analysis revealed that PNFI value is .583 so SAS has good fit of model. Although there are other authors who conclude that with sample size N=200 a reasonable estimate could be obtained for CFI and TLI, still researchers are warned since decisions strictly based on CFI could also cause wrong decisions because it also depends on sample size and hence rejection of fit model. For example, a correct model simulated with a sample N=200 and CFI value turned out to be .611 (poorly fitting model). It is noted that even with a relatively large sample size (N=500) a conventional cut off value of TLI may cause a correct model to be rejected (Shi, Lee, Maydeu-Olivares, 2018). Hu & Bentler (1999) notes sample size ≤ 250 could cause problems in Maximum Likelihood (ML) analysis. Thus, warns researchers to be cautionary on evaluation on model fit evaluation. Questions (items) assessing the same target or different items having nearly same meaning with different words may be the cause of correlated errors which in fact, may cause the correlate error terms (Meyer, n.d.). Since SAS has correlated error terms, this also concludes the idea that model could provide a multidimensional aspect if provided with enough number of items targeting the domain of interest. Be that as it may, Ellis (2017) states that if p value is < 0.05 and 0.05 < RMSEA < 0.08 then, null hypothesis is not exactly true but model has acceptable fit. Although generally accepted indice values are ≥ 0.80 for fit indices, a proposed common guideline for indice values follows as; very good fit ≥ 0.90; adequate but marginal fit ≥ 0.80-0.89; poor fit ≥ 0.60-0.79; very poor fit > 0.60 (Planing, 2014).

As a final thought, it is believed that the developed SAS might be used in different regions/states to compare the results and validate its purpose. For that reason, SAS with different samples is welcome. With this aim, both the created SAS in the original language and an English version of SAS are given in the appendixes.

ORCID

Adem AKKUŞ https://orcid.org/0000-0001-9570-3582

5. REFERENCES

Adesoji, F. A. (2008). Managing students' attitude towards science through problem–solving instructional strategy. *The Anthropologist*, *10* (1), 21-24.

Ajzen, I. (2005a). Behavioral interventions based on the theory of planned behavior: Brief description of the theory of planned behavior. Retrieved from http://people.umass.edu/aizen/pdf/tpb.intervention.pdf (accessed on 5 February 2019)

Ajzen, I. (2005b). Constructing a theory of planned behavior questionnaire: Brief description of the theory of planned behavior. Retrieved from http://people.umass.edu/aizen/pdf/tpb.measurement.pdf (accessed on 5 February 2019)

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. DOI: 10.1007/bf02294170

Anthoine, E., Moret, L., Regnault, A., Sébille, V., & Hardouin, J.-B. (2014). Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. *Health and Quality of Life Outcomes*, *12*, 2. DOI:10.1186/s12955-014-0176-2

Bernardez, R.Q. (1982). *Factors affecting attitudes to laboratory work*. Unpublished Master Thesis, Saint Louis University, Baguio, Philippines. [Abstract]

Brinkman, W-P. (2009). *Design of a questionnaire instrument, handbook of mobile technology research methods*. ISBN 978-1-60692-767-0, pp. 31-57 Netherlands: Nova Publisher

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

Byrne, B. M. &Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology, 30,* 557 - 576. DOI: https://doi.org/10.1177/0022022199030005001

Cabrera-Nguyen, P. (2010). Author guidelines for reporting scale development and validation results in the Journal of the Society for Social Work and Research. *Journal of the Society for Social Work and Research, 1*(2), 99-103.

Can, M., & Şahin, Ç. (2015). Okul öncesi öğretmen adaylarının fene ve fen öğretimine yönelik tutumlarının incelenmesi [Investigating Prospective Kindergarten Teachers' Science and Science Teaching Attitudes]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 15* (2), 13-26. DOI: 10.17240/aibuefd.2015.15.2-5000161311

Carlback, J. & Wong, A. (2018). *A study on factors influencing acceptance of using mobile electronic identification applications in Sweden*. Retrieved from http://www.diva-portal.org/smash/get/diva2:1214313/FULLTEXT01.pdf (accessed on 03 April 2019)

Cheung, D. (2007). Confirmatory factor analysis of the attitude toward chemistry lessons scale. Paper presented at *2nd NICE symposium*, Taipei, Taiwan, July 30-31, 2007.

Chinda, T., Techapreechawong, S., & Teeraprasert, S. (2012). *An investigation of relationships between employees' safety and productivity*. Retrieved from http://www.ppml.url.tw/EPPM/conferences/2012/download/SESSON4_A/10%20E145.pdf (accessed on 12 October, 2019)

Coll, R.L., Dalgety, J. & Salter, D. (2002). The development of the chemistry attitudes and experiences questionnaire (CAEQ). *Chemistry Education Research and Practice in Europe, 3*(1), 19-32.

Demirbaş, M. (2009). The relationships between the scientist perception and scientific attitudes of science teacher candidates in Turkey: A case study. *Scientific Research and Essays, 4*(6), 565-576.

Deshpande, L. (2004). Challenges in measurement of scientific attitude. Paper presented at *epiSTEME-1: An International Conference to Review Research on Science Technology and Mathematics Education* (137-138), Goa, India, December 13-17, 2004.

Dönmez, F., & Azizoğlu, N. (2010). Investigation of the students science process skill levels in vocational schools: a case of Balıkesir. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education, 4* (2), 79-109.

Ellis, J.E. (2017). Factor analysis and item analysis. *Applying Statistics in Behavioural Research* (pp. 11-59). Retrieved from https://www.applyingstatisticsinbehaviouralresearch.com/documenten/factor_analysis_and_item_analysis_version_11_.pdf (accessed on 22 December 2018)

Evrekli, E., İnel, D., Balım, A. G., & Kesercioğlu, T. (2009). The attitude scale of constructivist approach for prospective science teachers: a study of validity and reliability. *Journal of Turkish Science Education, 6*(2), 134-148.

Field, A. (2013). *Discovering statistics using ibm spss statistics* (4th Edition). London: SAGE

Finn, A.N. (2012) Teacher use of prosocial and antisocial power bases and students' perceived instructor understanding and misunderstanding in the college classroom. *Communication Education, 61*(1), 67-79, DOI: 10.1080/03634523.2011.636450

Foley, B., & McPhee, C. (2008). Students' attitudes towards science in classes using hands-on or textbook based curriculum. *AERA*, 1-12.

Fraser, B.J. (1981). *Test of science-related attitudes (TOSRA) handbook*. Victoria: Allanby

Garson, G.D., 2006. *Structural equation modelling.* North Carolina: G. David Garson and Statistical Associates Publishing

Hayton, J.C., Allen, D.G & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191-205. DOI: 10.1177/1094428104263675

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 2* (1), 104-121. DOI:10.1177/109442819800100106

Hinkin, T. R., Tracey, J. B., & Enz, C. A. (1997). Scale construction: Developing reliable and valid measurement instruments. *Journal of Hospitality & Tourism Research, 21*(1), 100-120. DOI:10.1177/109634809702100108

Hof, M.W. (2012). *Questionnaire Evaluation with Factor Analysis and Cronbach' s Alpha: An Example*. Retrieved from http://www.let.rug.nl/nerbonne/teach/rema-stats-meth-seminar/student-papers/MHof-QuestionnaireEvaluation-2012-Cronbach-FactAnalysis.pdf (accessed on 02 May 2016).

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. DOI: 10.21427/D7CF7R

Hu, L.-T., & Bentler, P. M. (1995). *Evaluating model fit*: in Structural Equation Modeling Ed. Rick H. Hoyle. London: Sage Publications

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. DOI: https://dx.doi.org/10.1080/10705519909540118

Johanson, G.A., & Brooks, G.P. (2010). Initial scale development: Sample size for pilot studies. *Educational and Psychological Measurement, 70*(3), 394-400. DOI: 10.1177/0013164409355692

Johnson, R.L. & Morgan, G.B. (2016). *Survey scales: Investigating scale quality*. New York, NY: The Guilford Press.

Kalaycı, Ş. (2010). Spss uygulamalı çok değişkenli istatistik teknikleri. (5. Baskı) [SPSS applied various statistical techniques (5th Edition)]. Ankara: Asil Yayın Dağıtım Ltd. Şti.

Kaya, Ç. & Altinkurt, Y. (2018). Öğretmenlerin psikolojik sermayeleri ile tükenmişlik düzeyleri arasındaki ilişkide psikolojik ve yapısal güçlendirmenin rolü [Role of Psychological and Structural Empowerment in the Relationship between Teachers' Psychological Capital and Their Levels of Burnout]. *Eğitim ve Bilim, 43* (193), 63-78, DOI: http://dx.doi.org/10.15390/EB.2018.6961

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd edition) New York: The Guilford Press

Knekta, E., Runyon, C., & Eddy, S. (2019). One Size Doesn't Fit All: Using Factor Analysis to Gather Validity Evidence When Using Surveys in Your Research. *CBE—Life Sciences Education, 18* (1), 1-17. DOI: https://doi.org/10.1187/cbe.18-04-0064

Kurnaz, M.A. & Yigit, N. (2010). Physics attitude scale: Development, validity and reliability. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education, 4* (1), 29-49.

Korkmaz, Ö., Şahin, A. & Yeşil, R. (2011). Bilimsel araştırmaya yönelik tutum ölçeği geçerlik ve güvenirlik çalışması [Study of Validity and Reliability of Scale of Attitude towards Scientific Research]. *Elementary Education Online, 10* (3), 961-973.

Lovelace, M. & Brickman, P. (2013). Best practices for measuring students' attitudes toward learning science. *CBE-Life Sciences Education, 12*(4), 606-617. DOI: 10.1187/cbe.12-11-0197

Meyer, J. (n.d.). Correlated errors in confirmatory factor analysis. Retrieved from https://www.theanalysisfactor.com/correlated-errors-in-confirmatory-factor-analysis (accessed on 17 September 2019).

Moore, R.W. & Foy, R.L.H. (1997). The scientific attitude inventory: A revision (SAI II). *Journal of Research in Science Teaching, 34*(4), 327-336.

Nayir, F. (2013). "Algılanan örgütsel destek ölçeğinin" kısa form geçerlik güvenirlik çalışması ["Perceived Organizational Support Scale"- Short Form Validity-Reliability Study]. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, 28, 89-106.

Nuhoglu, H. (2008). The development of an attitude scale for science and technology course. *Elementary Education Online*, *7*(3), 627-639.

Pedroso, R., Zanetello, L., Guimaraes, L., Pettenon, M., Goncalves, V., Scherer, J., Kessler, F., & Pechansky, F. (2016). Confirmatory factor anlaysis (CFA) of the crack use relapse scale (CURS). *Archives of Clinical Psychiatry, 43* (3), 37-40.

Planing, P. (2014). Innovation Acceptance: The Case of Advanced Driver-Assistance Systems: *Quantitative research approach* (pp. 230-231). Stuttgart: Springer

Sica, C. & Ghisi, M. (2007). The Italian versions of the Beck Anxiety Inventory and the Beck Depression Inventory-II: Psychometric properties and discriminant power. In M.A. Lange (Ed.), *Leading - Edge Psychological Tests and Testing Research* (pp. 27-50). New York: Nova

Schrodt, P., Witt, P.L., Turman, P.D., Myers, S.A., Barton, M.H & Jernberg, K.A. (2009) Instructor credibility as a mediator of instructors' prosocial communication behaviors and students' learning outcomes. *Communication Education*, *58*(3), 350-371, DOI: 10.1080/03634520902926851

Serin, O., & Mohammadzadeh, B. (2008). The relationship between primary school students' attitudes towards science and their science achievement (sampling: Izmir). *Journal of Educational Sciences, 2* (6), 68-75.

Shadfar, M. & Malekmohammadi, I. (2013). Application of Structural Equation Modeling (SEM) in restructuring state intervention strategies toward paddy production development. *International Journal of Academic Research in Business and Social Sciences, 3* (12), 576-618. DOI: 10.6007/IJARBSS/v3-i12/472

Shah, Z.A. & Mahmood, N. (2011). Developing a Scale to Measure Attitude towards Science Learning among School Students. *Bulletin of Education and Research, 33* (1), 71-81.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2018). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*(2), 310-334. DOI: https://doi.org/10.1177%2F0013164418783530

Tortop, H. S. (2013). Bilimsel Alan Gezisi Tutum Ölçeği Adaptasyon Çalışması [Adaptation Study of Attitude Scale towards Scientific Field Trips]. *Bartın Üniversitesi Eğitim Fakültesi Dergisi*, *2*(1), 228.

TTK. (2017, July 18). *Müfredatta yenileme ve değişiklik çalışmalarımız üzerine [On curriculum revision and changes work*]. Retrieved from https://ttkb.meb.gov.tr/meb_iys_dosyalar/2017_07/18160003_basin_aciklamasi-program.pdf (accessed on 25 August 2018).

Watkins, M. W. (2000). *Monte carlo PCA for parallel analysis [computer software]*. State College, PA: Ed & Psych Associates.

Wong., M. & Lian, S. (2003). Development of a self-efficacy scale for assessing secondary school students' science self-efficacy beliefs. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.472.8479 (accessed on 19 October 2014)

Vassallo, M., & Saba, A. (2015). Does money for grocery expenditure sway Italian consumers' motivational values in predicting Attitude towards eco-sustainable food products? *Contemporary Management Research, 11*(1), 3-22. DOI: doi:10.7903/cmr.13840

6. APPENDIX

Table A1. Bilimsel Tutum Ölçeği

Madde	Bilimsel Tutum	Kesinlikle Katılmıyorum	Katılmıyorum	Kararsızım	Katılıyorum	Kesinlikle Katılıyorum
1	Bilimsel gerçekler değişmez	1	2	3	4	5
2	Bilim adamları gerçeğin/fikirlerin değişebileceğine inanırlar/kabul ederler	1	2	3	4	5
3	Bilim adamları sorularının cevaplarını her zaman bulamazlar	1	2	3	4	5
4	Bilim adamları birbirlerinin çalışmalarını eleştirmemelidirler	1	2	3	4	5
5	Bilimsel çalışmalar bilim adamları içindir	1	2	3	4	5
6	Dinle çatışan konular çalışılmamalı/önemsenmemeli	1	2	3	4	5
7	Bir bilim adamının sahip olduğu en önemli araç hisleridir	1	2	3	4	5
8	Bilimsel gelişmeler daha sağlıklı yaşam sürmemizi sağlar	1	2	3	4	5
9	Deney yapmak derslerden daha zevklidir	1	2	3	4	5
10	Bilimsel keşifler faydadan çok zarar veriyor	1	2	3	4	5
11	Hocanın anlatmasındansa deney yaparak gerçekleri bulmayı tercih ederim	1	2	3	4	5
12	Farklı fikirleri hoş karşılarım	1	2	3	4	5
13	Fen dersleri zaman kaybıdır	1	2	3	4	5
14	Fen konuları zevksizdir	1	2	3	4	5
15	Deney yapmaktansa teorik bilgiler daha faydalıdır	1	2	3	4	5
16	Fen deneylerine daha az vakit verilmeli	1	2	3	4	5
17	Deneyler grup çalışmasıyla daha zevkli geçer/geçiyor	1	2	3	4	5
18	Bilimin temel amaçlarından biri yeni ilaçlar ve tedaviler bulmaktır	1	2	3	4	5
19	Yaşamı etkilediğinden İnsanlar bilimsel gerçekleri anlamalı	1	2	3	4	5
20	Bilim bir şeyin nasıl olduğunu açıklamaya çalışmaktır	1	2	3	4	5
21	Bilimsel çalışma bana zor gelir	1	2	3	4	5

Table A2. Scientific Attitude Scale

Item	Scientific Attitude	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
1	Scientific facts do not change	**1**	**2**	**3**	**4**	**5**
2	Scientists acknowledge/accept that facts may change	**1**	**2**	**3**	**4**	**5**
3	Scientists cannot always find the answers	**1**	**2**	**3**	**4**	**5**
4	Scientists should not criticize each other's work	**1**	**2**	**3**	**4**	**5**
5	Scientific works are for scientists	**1**	**2**	**3**	**4**	**5**
6	Topics contradicting with religion should not be studied/cared	**1**	**2**	**3**	**4**	**5**
7	The most important tool for a scientist is her/his feelings	**1**	**2**	**3**	**4**	**5**
8	Scientific progress helps us to have more healthy life	**1**	**2**	**3**	**4**	**5**
9	Doing experiments is more fun than having lectures	**1**	**2**	**3**	**4**	**5**
10	Scientific progress outputs harm more than good	**1**	**2**	**3**	**4**	**5**
11	I prefer to find facts rather than told by the instructor	**1**	**2**	**3**	**4**	**5**
12	I welcome different ideas	**1**	**2**	**3**	**4**	**5**
13	Science courses are waste of time	**1**	**2**	**3**	**4**	**5**
14	Science courses are tasteless	**1**	**2**	**3**	**4**	**5**
15	Theoretical knowledge is more helpful than experimenting	**1**	**2**	**3**	**4**	**5**
16	Science course hours must be reduced	**1**	**2**	**3**	**4**	**5**
17	Experiments are more fun with group works	**1**	**2**	**3**	**4**	**5**
18	One of the main aims of science is to find new cures	**1**	**2**	**3**	**4**	**5**
19	People should understand scientific facts since it affects life	**1**	**2**	**3**	**4**	**5**
20	Science is trying to explain things	**1**	**2**	**3**	**4**	**5**
21	Scientific works are baffling for me	**1**	**2**	**3**	**4**	**5**