IJATE

Volume 6 Issue 3 2019

International Journal of
Assessment Tools in Education

International Journal of Assessment Tools in Education

International Journal of
Assessment Tools in Education

http://ijate.net/

e-ISSN: 2148-7456

IJATE

e-ISSN 2148-7456

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

Volume 6 Issue 3 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/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. 3, 330–343
https://dx.doi.org/10.21449/ijate.546709

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

On the Statistical and Heuristic Difficulty Estimates of a High Stakes Test in Iran

Ali Darabi Bazvand [1,*], Shiela Kheirzadeh [2], Alireza Ahmadi [3]

[1] University of Human Development, College of Languages, Department of English Language, Sulaimani, Iraq
[2] Sobhe-Sadegh Institute of Higher Education, Isfahan, Iran
[3] Shiraz University, Faculty of Literature and Humanities, Department of English Language, Shiraz, Iran

ARTICLE HISTORY

Received: 29 March 2019
Revised: 12 May 2019
Accepted: 09 July 2019

KEYWORDS

Classical true score theory,
Heuristic difficulty,
High stakes test,
Item response theory,
Statistical difficulty

Abstract: The findings of previous research into the compatibility of stakeholders' perceptions with statistical estimations of item difficulty are not seemingly consistent. Furthermore, most research shows that teachers' estimation of item difficulty is not reliable since they tend to overestimate the difficulty of easy items and underestimate the difficulty of difficult items. Therefore, the present study aims to analyze a high stakes test in terms of heuristic (test takers' standpoint) and statistical difficulty (CTT and IRT) and investigate the extent to which the findings from the two perspectives converge. Results indicate that, 1) the whole test along with its sub-tests is difficult which might lead to test invalidity; 2) the respondents' ratings of the total test in terms of difficulty level are almost convergent with the difficulty values indicated by IRT and CTT, except for the two subtests where students underestimated the difficulty values, and 3) CTT difficulty estimates are convergent with IRT difficulty estimates. Therefore, it can be concluded that students' perceptions of item difficulty might be a better estimate of test difficulty and a combination of test takers' perceptions and statistical difficulty might provide a better picture of item difficulty in assessment contexts.

1. INTRODUCTION

To enhance the quality of educational systems, assessment is gradually taking the central role in the higher education process (Brown & Glasner, 1999). As a result, increasing attention has been paid to the academic standards with regard to the association between the students' entry level and the outcomes of the assessment (van de Watering & van der Rijt, 2006). However, as stated by van de Watering and van der Rijt (2006), "little is known about the degree to which assessments in higher education are correctly aimed at the students' levels of competence" (p. 134). This might have happened due to the obscured correspondence between test intentions and test effects (e.g., Cizek, 2012; Hubley & Zumbo, 2011; Xi, 2008) which might be associated with two technical expressions coined by Messick (1989), "construct-irrelevant variance" (CIV) and "construct underrepresentation". The former, which might be relevant to the present

CONTACT: Ali Darabi Bazvand ✉ alidarabi1350@gmail.com Lecturer of Applied Linguistics, University of Human Development, Faculty of Arts and Humanities, Department of English Language, Sulaimani, Iraq

ISSN-e: 2148-7456 /© IJATE 2019

study, occurs when the measure does not reflect the construct to be assessed; rather additional characteristics affect performance, while the latter happens when the measure fails to include important aspects of the construct (Cizek, 2012; Hubley & Zumbo, 2011; Knoch & Elder, 2013; Xi, 2008).

As one type of CIV, the difficulty level of the test items might affect test applicants' performance and hinder them from achieving the best level of their abilities. This would possibly make the test not to tap into the construct being measured and might render it unreliable and invalid. Therefore, research undertaken on item difficulty and the way teachers and students perceive item difficulty is germane to the assessment issues (van de Watering & Van der Rijt, 2006).

It is worth noting that the difficulty of an assessment instrument or items included in it might decrease the reliability of the assessment in two ways. First, if the difficulty level of the items was much higher than the students' ability level, this would result in loss of concentration, anxiety, decrease of motivation, confusion, uncertainty, etc. and as a consequence, more errors happen in assessment. Second, there is always the chance of guessing while answering test items, especially in multiple-choice tests. So, if the items are more difficult, this implies more students would guess and this allows more random errors to enter the variance of the assessment score (Bereby-Meijer, Meijer, & Flascher, 2002).

Moreover, in line with Messick's technical expressions of test invalidity, it is reported then that the difficulty level of test items seems to be an overriding factor in contributing to the test lapses and might create a mismatch between test score interpretation and test score use (e.g, Chappelle, Enright, & Jamison, 2010; Johnson & Riazi, 2013). Such a factor, more often than not, is considered to be a major cause for confusion, anxiety, uncertainty, and demotivation among test takers, and might subsequently motivate them to rely on guessing (Stanley, 1971).

In general, across the content of PhD entrance exams in Iran, it is assumed that such test lapses might exist which might be symptomatic of the test invalidity. Therefore, investigating the difficulty level of the test, by getting insight from stakeholders' perceptions (test takers' perspective in the case of the present study) and statistical quality of the test, as analyzed via CTT and IRT, is relevant. Considering the abovementioned points, the present study aimed at estimating the difficulty level of PhD Entrance Exam of ELT (PEEE, henceforth), a high stakes test in Iran, by taking both statistical and heuristic difficulty estimates, and whether the difficulty information yielded by both stakeholders' perception and statistical analyses converge.

1.1. Classical Test Theory (CTT)

Since the early 20th century, CTT has been used in estimating test/item difficulty. In relation to this theory, the knowledge/ability (represented by the true score of the test-takers) is defined as the expected score obtained by a student in a given test (Conejo, Guzmán, Perez-De-La-Cruz, & Barros, 2014).

The major assumptions underlying the CTT are: the mean of the test-takers' error score is zero; true scores and error scores are not correlated, and error scores obtained on the parallel tests are not correlated (Hambleton & Jones, 1993). According to Magno (2009), the assumption of classical test theory is that each test taker's score is a true score (unobservable) obtained if there were no errors in measurement. However, because the test instruments used are not perfect, the observed score of each test-taker might differ from his true ability.

In this theory, items are described by two parameters: the difficulty parameter, i.e., the proportion of the students who answered an item correctly, and the discrimination parameter, which will be estimated by the correlation between the item and the test score. As an early approach to estimate test/item difficulty, it suffers certain limitations such a considering all errors as random (Bachman, 1990); however, CTT is easy to use in several situations and it requires fewer number of testees, compared with other methods such as IRT.

1.2. Item Response Theory (IRT)

Item response theory is a probabilistic model that is to explain an individual's response to an item (Hambleton, Swaminathan, & Rogers, 1991). This theory is based on two main principles: (a) students' performance in a test would be explained by their level of knowledge, measured as an unknown numeric value h. (b) the students' performance estimated by the level of knowledge in answering an item would be probabilistically predicted and displayed using a function called the Item Characteristic Curve (ICC) (Hambleton, Swaminathan, & Rogers, 1991).

According to Birnbaum, (1968), there are three different models of IRT, namely one-parameter logistic model, two-parameter logistic model, and three-parameter logistics model. The one-parameter logistic model indicates the probability of a correct response as a logistic distribution where items differ merely regarding their difficulty and this model is used on multiple-choice (MC) or short response items which are dichotomous and do not allow for guessing (Birnbaum, 1968). The two-parameter logistic model, as stated by the same author, generalizes the one-parameter logistic model and allows items to differ not only regarding their difficulty but also differ in discriminating among individuals of various proficiency levels. Similar to the one-parameter logistic model, the two-parameter logistic model assumes that the probability of guessing is zero. Birnbaum also stated the three-parameter logistic model extends the two-parameter logistic model by including a guessing parameter which represents the probability of testees with low ability level correctly answer an item since for low ability testees, guessing is an influential factor in test performance.

To estimate item difficulty, the one-parameter IRT model using a single item parameter (i.e., difficulty parameter) is more frequently used (Van der Linden & Hambleton, 1997). The one-parameter model designates the probability of answering an item correctly through a logistic function indicating the difference between the proficiency level and the item difficulty. In justifying IRT use for difficulty estimation, Pardos and Heffernan (2011) stated, "Models like IRT that take into account item difficulty are strong at prediction" (p. 2). It should be mentioned that the one-parameter IRT model was used for the present study since the aim was merely estimating the difficulty of the items.

1.3. Local context

Since the evidence of item difficulty for the present study is provided by the PhD Entrance exam in Iran, it seems imperative to briefly introduce it here. High stakes tests in Iran have been considered as predominate tools to measure applicants' general and domain-specific knowledge and skills for the purpose of admission to higher education. Nevertheless, empirical studies have found that such tests, as levers of entering higher education, have fallen short of their expectations. That is, they have not been without their fair share of negative consequences (Farhady, 1998; Razmjoo, 2006). Specifically, findings from validity studies have shown that university entrance examinations in Iran are not socially responsive for graduate studies (Hajforoush, 2002; Shojaee & Gholipour, 2005).

As part of university entrance examinations, PhD entrance exams in Iran play a great role in the admission decisions of postgraduate studies. These high-stakes exams consist of a series of centralized written exams designed to screen PhD applicants (with different academic majors) to enter PhD programs. Since 2011, these exams superseded the traditional university-based examination sets in Iran. Administered by the National Organization for Educational Testing (NOET), they all appear in MC format with four-option items often consisting of three blocks: a general competence section, an academic talent test, and a field-specific section. For this study, the field-specific section of the PhD exam of ELT administered in 2014 was considered. More information on this exam is provided in the method section.

2. PREVIOUS STUDIES ON TEST DIFFICULTY

Previous research has highlighted various factors that might influence item difficulty, for instance, word knowledge (e.g., Rupp, Garcia, & Jamieson, 2001), negative stem (e.g., Hambleton & Jirka, 2006) and background knowledge of the topic (e.g., Freedle & Kostin, 1999). The purpose of conducting such studies was to make the item-writing process more efficient through academically publishing more detailed guidelines and item level descriptors to help item writers (Kostin, 2004). However, besides sensitivity to guidelines and item descriptors, in Bachman's (2002) words, "difficulty does not reside in the task alone but is relative to any given test-taker" (p. 462). Therefore, who would take the test and answer items would definitely influence the way items are designed and developed. Hambleton and Jirka (2006) recommended asking experts in the field of test development and scoring to estimate the task difficulty. However, even these experts are not necessarily accurate in their predictions of task difficulty in both first language (L1) (Bejar, 1983; Hambleton & Jirka, 2006) and second language (L2) tests (Bachman, 2002; Elder, Iwashita, & McNamara, 2002).

Bejar (1983) concluded that a group of four test developers could not make a reliable difficulty estimation for L1 writing tasks. As with L2 tests, Alderson (1993) reported that experienced item writers and raters were somewhat better than inexperienced ones on predicting item difficulty; however, the significance of this difference was not estimated. Hamp-Lyons and Mathias (1994) reported a considerable agreement between expert judges (two raters familiar with the test and two L2 writing experts); however, there was an astonishingly reverse relationship between the difficulty predicted by experts and raters and the actual difficulty of the test. Therefore, as suggested by Lee (1996), students might be able to estimate difficulty more accurately than teachers. Nevertheless, teachers/experts' estimation has received more attention than students' estimation.

Wauters, Desmet, and van Den Noortgate (2012, p. 1183) compared six different estimations of the difficulty: "proportion correct, learner feedback, expert rating, one-to-many comparison (learner), one-to-many comparison (expert) and the Elo rating system" with the IRT-based calibration. Results revealed that proportion correct showed the strongest relation with IRT-based difficulty estimates, followed by student estimation. The participants of the study included 13 teachers and 318 students (secondary level) in the field of Linguistic and Literature. The researchers concluded that student estimations were somewhat better. To explain the difference in the rating of the two groups of the participants, the researchers referred to the much larger sample size of the students, compared to the teachers.

In a more recent study, Conejo, Guzmán, Perez-De-La-Cruz, and Barros (2014, p. 594-595) named three test/task difficulty estimation approaches.

- Statistical, that is, estimating the difficulty from a previous sample of students.
- Heuristic, that is, by human ''experts'' direct estimation.
- Mathematical, given a formula that predicts the difficulty in terms of the number and type of concepts involved in the task

To estimate difficulty using statistical approaches, the definition of the concept of difficulty need to exist. Therefore, this approach is commonly associated with using CTT or IRT in the assessment. From the heuristic standpoint, teachers or course designers are commonly experts that estimate the difficulty; however, students might also be considered as experts in this approach. In mathematical approaches, difficulty would be estimated by a formula that uses a number of item/task features, e.g., complexity or the number of concepts involved. As such, the focus of the present study is to estimate the statistical (CTT and IRT estimations) and heuristic (test-takers' standpoint) difficulty of test items and investigate the extent to which findings from the two perspectives are congruent.

3. METHOD

3.1. Participants

The participants in the current study included PhD applicants and first semester PhD candidates of Iran majoring in ELT. Test score data for a population of 999 PhD exam applicants (397 females and 602 males) participating in January 2011 administration of this test was analyzed in terms of the difficulty level. Performance data for this population was provided by the National Organization for Educational Testing (NOET) at the request of Shiraz University, Iran. No information regarding their age, names, average score, and the socioeconomic status was provided by this organization.

The second group of participants was a sample of 103 PhD candidates of ELT who had been admitted to the PhD programs. Their ages ranged between 25 and 40, with 46 of them being female and 57 of them being male. They were recruited to respond to the survey questionnaires. Since it was not feasible to obtain a complete list of all the participants from whom to make a random selection, a snowball sampling procedure was preferred. This particular sample was targeted, since tracking them to administer the questionnaire was less likely to be problematic. In addition, they were in a better position to recollect their test-taking experience than those who had taken it earlier. They received the questionnaires through email. Upon their views, they provided evidence with regard to the characteristics of PEEE in terms of its difficulty level. A brief summary of the participants' self-reported background is provided in Table 1.

Table 1. Background Information Reported by PhD Students (n=103)

Variable	Level	F (%)
Gender	Male	57(55.3%)
	Female	46(44.7%)
	Total	103(100%)
Age	25-27	13(12.6%)
	28-30	38(36.9%)
	30-39	41(39.8%)
	40+	11(10.7%)
Times taking exam	First	10(9.7%)
	Second	60(58.3%)
	Third	24(23.3%)
	Fourth	9(8.7%)
Field-specific test scores	Less than 30%	5(4.9%)
	30-40%	31(30.1%)
	40-50%	48(46.6%)
	50+	42(40.8%)
General English test scores	Less than 30%	6(5.8%)
	30-40%	24(23.3%)
	40-50%	31(30.1%)
	50+	42(40.8%)

3.2. Instruments and Data Collection

Two types of instruments were used to collect the data for this study, namely PEEE test score data and PhD students' questionnaires. PEEE is a field-specific exam which is aimed at measuring the PhD candidates' expertise in the field of English Language Teaching (ELT) and is supposed to be related to the courses students have passed in the MA or even BA program.

In fact, it assesses the students' domain-specific knowledge in areas which are the prerequisite for entering the PhD programs since the PhD program is built on such areas of knowledge. It consists of 100 items including questions on Linguistics (15 items), Teaching Methodology (15 items), Research Methods (15 items), Language Testing and assessment (15 items), Theories of SLA (30 items), and finally Discourse & Sociolinguistics (10 items).

PhD students' questionnaire comprised of 24 items and categorized into two parts to provide information on students' background and their perceptions with regard to test characteristics. For test characteristics part, the options included *very difficult*, *difficult*, *average*, *easy* and *very easy*. The reliability of the whole questionnaire was reported to be .73, as estimated through Cronbach's alpha. The validity of questionnaires was established using expert judgment.

3.3. Data Analysis

For the data analysis, both questionnaire and test score data were analyzed. With regard to the questionnaire, stakeholders' perceptions were analyzed for the difficulty level of the test. For this reason, a series of Binomial tests of significance were used to report the participants' responses to the specified questionnaire items in the form of observed proportions. Concerning the PEEE test score data, CTT, IRT and Cronbach's alpha were applied to estimate the difficulty level and the reliability coefficients of the whole test and its subtests, respectively.

4. RESULTS

For investigating the difficulty level of the test, the study benefitted from heuristic analysis, i.e., stakeholders' perceptions (via questionnaire) and statistical analysis, i.e., CTT and IRT analysis. The details are explained below.

4.1. Heuristic difficulty of the items

PhD students' responses to questionnaires revealed some important findings. They, almost all, did express the same collective opinion with regard to the level of difficulty of the items. As shown in Table 2, of 103 respondents, about half of them (58%, *p* =.114), answered that the total test is difficult. It is also reported that some subtests like Teaching Methodology (49%) and Linguistics (46.1%) designed based on the BA courses are moderately difficult and some others like Theories of SLA (64%), Language Testing and assessment (73%), and Discourse & Sociolinguistics (63%) which were based on MA courses are reported to be significantly difficult.

Table 2. Stakeholders' Perceptions of Difficulty of PEEE

PEEE and its Subtests	Category	N	Observed Prop.	Test Prop.	Sig. (2-tailed)
Total test	Easy*	43	.42	.50	.114
	Difficult +	60	**.58**		
Linguistics	Easy*	55	.54	.50	.000
	Difficult+	48	**.46**		
Teaching Methodology	Easy*	52	.51	.50	000
	Difficult+	51	**.49**		
Theories of SLA	Easy*	37	.36	.50	.006
	Difficult+	66	**.64**		
Language Testing and assessment	Easy*	28	.27	.50	.000
	Difficult+	75	**.73**		
Research Methods	Easy *	44	.43	.50	.001
	Difficult+	59	**.57**		
Discourse & Sociolinguistics	Easy*	38	.37	.50	.010
	Difficult+	65	**.63**		

* Combined 'Easy' and 'Very easy' responses

+ Combined 'Difficult' and 'Very difficult responses

4.2. Statistical difficulty of the items

4.2.1. *CTT difficulty*

In addition to the analysis of stakeholders' perceptions with regard to the level of difficulty, the test was also subjected to statistical item analysis. In this procedure, the difficulty index (referred to as a p-value) was estimated as the proportion of examinees correctly answering each item. As such, items shown to have demonstrated values above .80 or below .40 were considered to be too easy or too difficult, respectively (Apostolou, 2010); therefore, their difficulty level is not desired. With regard to the present study, all the subtests of PEEE were subjected to item analysis.

The first specialized subtest included in the PEEE was "Teaching Methodology" consisting of 15 items. As indicated by item analysis, the results from Table 3 reveal that the difficulty level of this subtest amounts to .39 with the difficulty values of individual items ranging from .52 to .07. As it is reported, of 15 items included in this subtest, 12 items do not fall within the above criterion range of difficulty, revealing that this subtest is somehow difficult.

Table 3. Difficulty Level of the Total Test and its Subtests

Subtest	Number of items	Mean Difficulty
Total Test	100	.24
Teaching Methodology	15	**.39**
Linguistics	15	.32
Research Methods	15	.25
Language Testing and assessment	15	.15
Language Skills	10	.21
Theories of SLA	20	.19
Discourse & Sociolinguistics	10	.23

The second subtest subjected to item analysis was "Linguistics" subsisting of 15 items. With regard to this subtest, Table 3 displays that the difficulty value of the whole subtest (p =.32) is not desired. Hence it provides evidence that this subtest is difficult.

The third subtest analyzed for difficulty index was "Research Methods". Like the first two sections, this subtest consists of 15 items. As Table 3 demonstrates, the difficulty value of the whole test is .25, falling far below the acceptable estimate of the desired difficulty. This finding is also true for individual items. Of the total of 15 items analyzed, 13 of them demonstrated difficulty values lower than the least acceptable criteria of the desired difficulty, suggesting that this subtest is also difficult.

The fourth subtest subjected to item analysis was "Language Testing and Assessment". Concerning this subtest, the results from Table 3 show that with a difficulty index of .15, this subtest might have been much too difficult for the applicants. Of particular interest is that no individual item displayed a difficulty value greater than the least desired difficulty of .40; such finding reveals that this subtest is problematic and might have introduced substantial CIV into the test scores.

The fifth subtest analyzed in terms of difficulty level was "Language Skills" consisting of 10 items. With regard to test difficulty, Table 3 shows a low index of difficulty (*p* =.21). As for individual items, it is reported that no items demonstrated a difficulty value more than the least desired yardstick (p =.40); therefore, introducing substantial CIV in the test scores.

The sixth subtest of PEEE analyzed for item difficulty was "Theories of SLA" subsisting of 20 items. As displayed in Table 3, the difficulty index reported for the whole subtest is .19. As for individual items, no difficulty value was reported to exceed the least acceptable criterion, showing that the test is unduly difficult.

The last area of investigation for item analysis was “Discourse & Sociolinguistics”, both of which being considered as one subtest and consisting of 10 items. As it is evident in Table 3, the estimated difficulty value reported for the whole test was .23 which was far too low as measured against the least desired yardstick. Like other subtests, in this section, the difficulty indices for all of the individual items were shown to be dramatically lower than the acceptable criteria, indicating that this subtest is also very difficult. In a nutshell, the overall results from the item analysis refer to the PEEE as being substantially difficult for PhD students.

4.2.2. *IRT difficulty*

In addition to the statistical analysis of CTT and stakeholders’ perceptions with regard to the level of difficulty, the test was also subjected to IRT analysis. In this procedure, the theoretical range of item difficulty falls within the range of -∞ to + ∞ on the ability scale, but in practice, the empirical range falls within the area of -2 to +2 (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). Items are shown to demonstrate *b* values (difficulty estimates) near -2 correspond to very easy items that are at the left or the lower end of the ability scale and items displaying *b* values near +2 are considered as very difficult that fall at the right or higher end of the ability scale. In order to have a better understanding, Baker (2001) defined the difficulty level of an item in verbal terms with their corresponding empirical ranges of *b* parameter as follows:

Table 4. Difficulty Parameter Values (from Baker, 2001, p. 12)

Verbal Label	Range of *b* values
Very easy	– 2.0 and below
Easy	– 2.0 ~ – 0.5
Medium	– 0.5 ~ + 0.5
Difficult	+ 0.5 ~ + 2.0
Very difficult	+ 2.0 and over

With regard to the present study, all the subtests of PEEE were subjected to IRT difficulty analysis. In the interest of brevity, only the difficulty values for the overall test as well as subtests are presented here. As indicated by test difficulty analysis, results from Table 5 reveal that the difficulty level of Teaching Methodology subtest amounted to 5.18. Based on the yardstick reported in Table 4, this subtest was considered *very difficult* and among the 15 items included in this subtest, 12 items fell beyond + 2.0, as the criterion range of *b* value; that is, they were *very difficult* and the remaining 3 items fell within the range of + 0.5 ~ + 2.0, being considered as *difficult*. Worthy of note is that the difficulty value for some items amounted to 10, showing that they were much beyond the ability level of examinees.

The second subtest subjected to *b* parameter analysis was Linguistics subsisting of 15 items. With regard to this subtest, Table 5 displays that the difficulty value of the whole test (*b* = 4.05) which was beyond + 2.0, demonstrated that it was *very difficult*. Regarding the individual items, 10 items were considered as *very difficult*, 2 as *difficult*, one item as *medium*, and 2 items as *easy*.

The third subtest analyzed for difficulty index was Research Methods. Like the first two sections, this subtest consisted of 15 items. As Table 5 demonstrates, the difficulty value of the whole test fell beyond + 2. This finding was also true for most of the individual items. Of the total of 15 items analyzed, 14 of them demonstrated difficulty values beyond + 2.0, and one item fell within the range of *difficult* items.

Table 5. Results of Test Difficulty Parameter in IRT Model

Subtest	Mean Difficulty	SD
Total Test	3.86*	1.70
Teaching Methodology	5.18*	2.87
Linguistics	4.05*	3.27
Research Methods	3.90*	.52
Language Testing and assessment	3.45*	.93
Language Skills	3.87*	1.30
Theories of SLA	3.41*	1.34
Discourse & Sociolinguistics	3.13*	.99

* Larger than + 2.0. (Very difficult)

The fourth subtest subjected to item analysis was Language Testing and Assessment. Concerning this subtest, the results from Table 5 show that with a b value of 3.45, this subtest was *very difficult* for the applicants. Of particular interest was that 14 items display a *b* value greater than the least value for *very difficult* items and one item covered the range of *difficult* items; this finding reveals that this subtest was substantially difficult.

The fifth subtest analyzed in terms of difficulty level was Language Skills consisting of 10 items. With regard to test difficulty, Table 5 shows a greater index of difficulty (b = 3.87) than + 2.0. As for individual items, it was found that 9 out of 10 items demonstrated a difficulty value more than the yardstick for *very difficult* items. Only one item was reported as difficult.

The sixth subtest of PEEE analyzed for test difficulty was Theories of SLA subsisting of 20 items. As displayed in Table 5, the difficulty *b* parameter reported for it was 3.41, symptomatic of *very difficult* tests. With regard to the individual items, it was found that 17 items displayed difficulty values larger than + 2.0, suggesting that they were *very difficult*, with the remaining 3 items fell under the category of *difficult* items.

The last area of investigation for item analysis is Discourse and Sociolinguistics, both of them were considered as one subtest and consisted of 10 items. As it is evident in Table 5 above, the estimated *b* value reported for this subtest was 3.13, indicating the test was *very difficult*, as measured against the yardstick of + 2.0. Like other subtests, in this section, the difficulty indices for almost all of the individual items were shown to be dramatically larger than the yardstick labeled for *very difficult* items.

Finally, as indicated in Table 5 as well as in Figure 1, the total test was shown to be *very difficult (*3.86*)*. As such, it can be argued that, based on the results from the IRT difficulty analysis, the PEEE test is prone to unreliability.

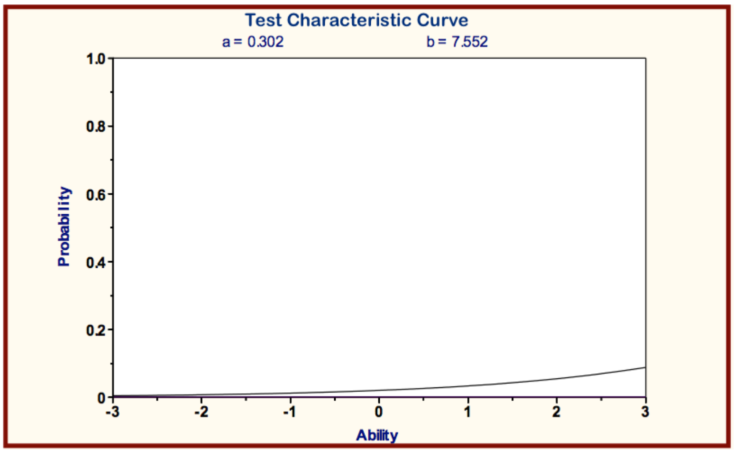


Figure 1. Total Test Difficulty: Test Characteristic Curve

4.2.3. *Comparison between heuristic and statistical difficulty*

As demonstrated in Table 6 below, results revealed that the respondents' rating of the total test in terms of difficulty level was almost convergent with the difficulty values indicated by IRT and CTT difficulty analyses, with reference to the same subtests. However, there were some specific cases of inconsistency between the results from heuristic difficulty and statistical difficulty; the results reported for the heuristic difficulty showed moderate difficulty values for Linguistics and Teaching Methodology subtests, while the findings from IRT and CTT difficulty demonstrate *very difficult* description for the same subtests. To recapitulate, when comparing the heuristic and statistical results for difficulty level, most of the subtests in the heuristic difficulty classification displayed the label of *difficult* and *very difficult*, while most of the subtests in the statistical category demonstrated the label of *very difficult*. This finding might lead to the overall conclusion that PEEE is a difficult test. As such, inappropriate test difficulty level was considered as evidence for invalidity of PEEE.

Table 6. Comparison between Heuristic and Statistical Difficulty

Test	Heuristic Difficulty	Statistical Difficulty	
	Questionnaire	CTT	IRT
Total Test	Difficult (58 %)	Very difficult (.24*)	Very difficult (3.86*)
Teaching Methodology	Moderate (49 %)	Difficult (.39*)	Very difficult (5.18*)
Linguistics	Moderate (46 %)	Difficult (.32*)	Very difficult (4.05*)
Research Methods	Difficult (57 %*)	Very difficult (.25*)	Very difficult (3.90*)
Language Testing	Very difficult (64 %*)	Very difficult (.15*)	Very difficult (3.45*)
Language Skills	Very difficult (64 %*)	Very difficult (.21*)	Very difficult (3.87*)
Theories of SLA	Very difficult (63 %*)	Very difficult (.19*)	Very difficult (3.41*)
Discourse & Sociolinguistics	Very difficult (73 %*)	Very difficult (.23*)	Very difficult (3.13*)

* Heuristic difficulty values above % 50 (difficult & very difficult)
* CTT difficulty values below 0 .40 (difficult & very difficult)
* IRT difficulty values larger than + 2.0 (very difficult)

5. DISCUSSION and CONCLUSION

The present study investigated the statistical and heuristic difficulty of PEEE in Iran. Findings of the study demonstrated that the statistical and heuristic difficulty investigations converge, indicating that the PEEE test is unduly difficult for test applicants. Results of the analysis of questionnaire items showed that for most of the PhD students (58%), the total test was very difficult.

IRT analysis of test difficulty also showed that all subtests were labeled as *very difficult* as compared with the criterion (+2.0 and beyond for very difficult items) recommended by researchers (Baker, 2001). Specifically, some items displayed values as large as 9.0, suggesting that they were much beyond the ability level of test applicants. The overall results from the *b* parameter IRT analysis of the PEEE subtests, and in most cases, their individual items together with the results from stakeholders' perceptions denote the PEEE test as very difficult. This finding can be regarded as good evidence for invalidity of this test (at least in terms of difficulty level). Moreover, the findings from the comparison between statistical and heuristic analysis showed that they were almost convergent, though there were some minor contradictions. One possible explanation might rest on the fact that, for the main part, the content of PEEE test is not based on the courses PhD applicants have passed but on those sources that they are not aware of or at least a few applicants have the chance to make use of. This could make the test difficult and might systematically introduce CIV into observed scores.

In a similar study, Rezvani and Sayyadi (2016) investigated the validity of new Iranian TEFL

PhD program entrance exam by asking PhD instructors and students. The result of their study revealed that, “the new exam was perceived to demonstrate defective face, content, predictive, and construct validities” (p. 1111). Razavipur (2014) studied the substantive and predictive validity facets of the university entrance exam for English majors by asking the ideas of 111 English major university students. He found that a large number of construct-irrelevant items exist in the exam along with a number of items that make no unique contribution to the exam. Furthermore, this finding was supported by research, though on a different testing application context. For example, in Apostolo's (2010) study, candidates’ heuristic task difficulty in the KPG listening tests was found to correlate to a great extent with the results of item analysis.

The findings of the present study might be somewhat consistent with Hamp-Lyons and Mathias (1994) who reported an astonishingly reverse relationship between the difficulty predicted by experts and raters and the actual difficulty of the test. As it was the case in the present study, the present so-called standard exam turned out to be a highly difficult one both heuristically and statistically. In the words of Nickerson (1999), when one decides to assess others’ knowledge and information, he requires to make a mental model of what they might know and if he has no access to specific information regarding those target group, a faulty mental model would be formed.

As such, any indiscriminate dealing with these tests regarding their interpretation and use might generate negative impacts on different stakeholders, across different testing contexts. Therefore, test practitioners should exercise high care when dealing with these gatekeeping tests in terms of item writing, test construction and test administration and also, as stated by Bachman (2002), difficulty is not just due to the tasks but it is a relative concept that varies across test-takers. As stated by Elder, Iwashita, and McNamara (2002, p. 350),

> *If test-takers can predict what makes a task difficult, it may be wise for us to access their views during the test design stage to determine whether they correspond to the hunches of test-developers and with existing theories about what makes a task more or less complex. It is conceivable that test-takers may be able to identify additional features of the task, or additional challenges involved in performing such tasks other than those visible to the test-developer or to the rater.*

Finally, the findings might be discussed from the social projection perspective, i.e., ascribing, generalizing and projecting what we know (the item developers in the case of the present study) to others (test-takers). In this regard, Nickerson (1999) stated that high familiarity with the particular topic might lead to over-ascription of what one knows to others. Also as stated by Goodwin (1999), judges (or item designers as it is the case in the present study) are typically experts in their fields. Since they might be much more knowledgeable in the related field, they might not be able to put themselves in the place of students adequately. Furthermore, their expectations of the examinees are possibly too high and they might also have difficulty differing between the proportion of examinees who should have answered an item correctly and who could have answered an item incorrectly.

In other words, the item writers might differ in their backgrounds and levels of experience with students. Item writers might tend to overestimate the performance of students. They might have based their judgments on “what they think students ought to know” (Verhoeven et al., 2002, p. 865). Such claim was supported by Impara and Plake (1998) who stated that even though judges (the expert in the field who design items/tests) are in close contact with the educational program, there is still a large difference in cognitive levels between them and the students. As it might be the case in the present study, the designers of PEEE exam, due to their familiarity with the subject matter, overgeneralized it to the test takers whose result was a very difficult test from the perspective of test takers.

Therefore, as the focus of the present study was a high-stakes test, the designers of such tests are recommended to consider the learners' characteristics and various possible learning environments in mind while developing items since very difficult test/items result in loss of concentration, anxiety, decrease of motivation, confusion, and uncertainty on the side of the test-takers. Such implication is in line with Bachman (2000, as cited in Brindley & Slatyer, 2002) who stated that, as soon as one considers what makes items difficulty, one immediately realizes that difficulty is not a reasonable question at all. A given task or item is differentially difficult for different test takers and a given test taker will find different tasks differentially difficult. Ergo, difficulty is not a separate quality at all, but rather a function of the interaction between task characteristics and test taker characteristics. When we design a test, we can specify the task characteristics, and describe the characteristics of the test takers, but getting at the interaction is the rub. Therefore, future researchers are recommended to work on item-writing guidelines used by item writers to see if these guidelines match the expectations, needs, and requirement of the target populations taking the test, especially in the case of high-stakes tests.

ORCID

Ali Darabi Bazvand https://orcid.org/0000-0002-2620-4648
Shiela Kheirzadeh https://orcid.org/0000-0003-4665-0554

6. REFERENCES

Alderson, J. C. (1993). Judgments in language testing. In D. Douglas & C. Chapelle (eds.), *A new decade of language testing* (pp. 46–57). Arlington. VA: TESOL.

Apostolou, E. (2010). Comparing perceived and actual task and text difficulty in the assessment of listening comprehension. In *Lancaster University Postgraduate Conference in Linguistics & Language Teaching* (pp. 26-47).

Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453–476.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford university press.

Baker, F. (2001). *The basics of item response theory.*, College Park: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*, 303–310

Bereby-Meijer, Y., Meijer, J., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making*, *15*, 313–327.

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.* Reading, MA: Addison-Wesley.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*, 369-394.

Brown, S., & Glasner, A. (1999). *Assessment matters in higher education*. Buckingham: SRHE and Open University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference?. *Educational Measurement: Issues and Practice*, *29*, 3-13.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*, 31.

Conejo, R., Guzmán, E., Perez-De-La-Cruz, J. L., & Barros, B. (2014). An empirical study on the quantitative notion of task difficulty. *Expert Systems with Applications*, *41*, 594-606.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer?. *Language Testing*, *19*, 347-368.

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

Farhady, H. (1998). A critical review of the English section of the BA and MA University Entrance Examination. In the *Proceedings of the conference on MA tests in Iran*, Ministry of Culture and Higher Education, Center for Educational Evaluation. Tehran, Iran.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*, 2-32.

Goodwin, L. D. (1996). Focus on quantitative methods: Determining cut-off scores. *Research in Nursing & Health, 19*, 249–256.

Hajforoush, H. (2002). Negative consequences of entrance exams on instructional objectives and a proposal for removing them. Proceedings of *the Isfahan University Conference on Evaluating the Issues of the Entrance Exams*.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12,* 38-47.

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

Hambleton, R., & Jirka, S. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 399–420). Mahwah, NJ: Erlbaum.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, *3*, 49–68.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, *103*, 219.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69–81.

Johnson, R.C., & Riazi, M. (2013). Assessing the assessments: Using an argument-based validity framework to assess the validity and use of an English placement system in a foreign language context. *Papers in Language Testing and Assessment, 2*, 31-58

Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessments (PELAs). *Papers in Language Testing and Assessment*, *2*, 48-66.

Kostin, I. (2004). *Exploring item characteristics that are related to difficulty of TOEFL dialogue items* (TOEFL Research Rep. No. 79). Princeton, NJ: ETS.

Lee, F. L. (1996). *Electronic homework: an intelligent tutoring system in mathematics*. (Doctoral Dissertation). The Chinese University of Hong Kong. Hong Kong, China.

Lee, F. L., & Heyworth, R. M. (2000). Problem complexity: a measure of problem difficulty in algebra by using computer. *Education Journal, 28*, 85–107.

Magno, C. (2009). Demonstrating the difference between Classical Test Theory and Item Response Theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1,* 1-11.

Nickerson, R. S. (1999). How we know-and sometimes misjudge-what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*, 737–759.

Pardos, Z. A., & Heffernan, N. T. (2011). KT-IDEM: Introducing item difficulty to the knowledge tracing model. In J. Konstan, R. Conejo, J. L. Marzo, & N. Oliver (Eds.), Proceedings of the *19th international conference on user modeling, adaptation and personalization* (Vol. 6787, pp. 243–254). Lecture Notes in Computer Science.

Razavipur, K. (2014). On the substantive and predictive validity facets of the university entrance exam for English majors. *Research in Applied Linguistics*, *5*, 77-90.

Razmjoo, S. A. (2006). A content analysis of university entrance examination for English majors in 1382. *Journal of Social Sciences and Humanities, Shiraz University, 46,* 67-75.

Rezvani, R., & Sayyadi, A. (2016). Ph. D. instructors' and students' insights into the validity of the new Iranian TEFL Ph. D. program Entrance Exam. *Theory and Practice in Language Studies*, *6*, 1111-1120.

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, *1*, 185-216.

Shojaee, M. & Gholipoor, R. (2005). *Recommended draft of applying university student system survey and designing acceptance model of university student*. Research Center of the Parliamnet, No. 7624.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356-442). Washington, DC: American Council on Education

van de Watering, G., & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review, 1,* 133-147.

van der Linden, W., & Hambleton, R.K. (1996). Item response theory: Brief history, common models, and extensions. In W. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item-response theory* (pp. 1–28). Berlin: Springer-Verlag.

Verhoeven, B. H., Verwijnen, G. M., Muijtjens, A. M. M., Scherpbier, A. J. J. A., & Van der Vleuten, C. P. M. (2002). Panel expertise for an Angoff standard setting procedure in progress testing: Item writers compared to recently graduated students. *Medical Education, 36,* 860–867.

Wauters, K., Desmet, P., & van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education, 58*, 1183–1193.

Xi, X. (2008). Methods of test validation. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education, 2nd edn, vol. 7: Language testing and assessment* (pp. 177–196). New York: Springer.

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 344–361
https://dx.doi.org/10.21449/ijate.544277

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

Formative Assessment of Writing (FAoW): A Confirmatory Factor Structure Study

Elaheh Tavakoli [1], Mohammad Reza Amirian*[,1], Tony Burner [2], Mohammad Davoudi [1], Saeed Ghaniabadi [1]

[1] Department of English Languages and Literature, University of Hakim Sabzevari, Sabzevar, Iran
[2] University of South-Eastern Norway, Norway

ARTICLE HISTORY

Received: 24 March 2019
Revised: 21 May 2019
Accepted: 09 July 2019

KEYWORDS

Formative Assessment of Writing,
Factor Solution,
Formative Assessment,
Confirmatory Factor Analysis (CFA)

Abstract: This validation study was undertaken to evaluate the construct of Formative Assessment of Writing (FAoW) operationalized by an instrument with 50 Likert scale items. To identify the EFL learners' experiences of FAoW practices, the instrument was first piloted with three EFL learners, and subsequently administered on a sample of 255 EFL learners selected based on purposive sampling. A five-factor solution with five latent variables (i.e. clarifying criteria, evidence on students' current learning, feedback to move learners forward, peer-assessment and autonomy) was evaluated through Confirmatory Factor Analysis (CFA) with AMOS 22. Model fit showed that the five-factor structure of FAoW could only be supported in terms of absolute and parsimony fit indices. The model with three factors (i.e. clarifying criteria, peer-assessment and feedback) in two stages of pre- and while-writing, however, provided higher discriminant validity in addition to absolute and parsimony fit indices. In other words, FAoW was not found to be practiced within its full potential with five components in the context of this study. A conceptual model was developed based on the findings and the literature to show pedagogical application of FAoW and how it can be practiced in line with Black and Wiliam's (2009).

1. INTRODUCTION

Since Formative Assessment (FA) was introduced to the field of education in the late 1990s by the Assessment Reform Group in the UK (e.g. Black & Wiliam 1998), many scholars, particularly in Europe and the USA, tried to investigate its theoretical base and practice. Most importantly and more closely related to this study, Black and Wiliam (2006; 2009) tried to provide a unifying theoretical framework for FA practices after interviews with teachers who developed FA and observation of the changes that occurred in their classrooms.

In second language (L2) writing, however, FA has been underexplored and much of the available research has focused on summative assessment, peer assessment or the effectiveness

CONTACT: M. R. Amirian ✉ smr.amirian@hsu.ac.ir Department of English Languages and Literature, University of Hakim Sabzevari, Sabzevar, Iran

ISSN-e: 2148-7456 /© IJATE 2019

of teachers' feedback (Burner, 2015; Lee, 2003; 2011). Formative Assessment of Writing (FAoW) is a prospective and aims to improve learning and fill the gap between students' current and potential state of development. It is a construct which has not been adequately defined, operationalized and validated so far. This study is a response to Johnson and Riazi (2017), who referred to the lack of local validation efforts for ensuring that the writing instruments are compatible with the unique learning outcomes, students, and context of the program. Tavakoli, Amirian, Burner, Davoudi and Ghaniabadi (2018) developed a FAoW instrument (Appendix I) which consisted of a comprehensive list of FA practices in writing classrooms based on Black and Wiliam's (2009) FA framework and Hattie and Timperley's (2007) feedback model. This study is factor structure of that instrument and part of a PhD project to investigate FAoW from both teachers' and students' perspective. In the project, two parallel versions of FAoW instrument were developed: EFL students' experiences of their teachers' FAoW practices and EFL teachers' perspective about their own practice of FAoW. Our earlier article Tavakoli, et al., (2018) pertained to the theoretical foundation and the development of FAoW instruments. This research is an attempt to validate the students' version through CFA. In this study, the words item, experience and practice are used interchangeably as every item in FAoW instrument is a teacher's FA practice or classroom activity which the students reported the frequency of their experience.

1.2. Review of the literature on FAoW

The literature on FAoW has highlighted some studies (e.g. Burner, 2015; 2016; Lee, 2007; 2011; Lee & Coniam, 2013; Mak & Lee, 2014; Naghdipour, 2016; 2017; Saliu-Abdulahi, 2017; Saliu-Abdulahi, Hellekjær, & Hertzberg, 2017; Tavakoli, et al,, 2018; Wingate 2010). The construct of FA in general has been described and conceptualized in various ways (Bennett, 2011); different scholars have developed different writing assessment instruments which could be used formatively. However, "there is no one definition of formative assessment of writing" (Burner, 2016, p. 4).

In the current research, we probed into FAoW considering FA model in general and assessment practices and feedback on the students' writing assignments in particular. The construct of FAoW has been operationalized in some studies which are worth citing here. In line with the ten principles of FA (aka Assessment for Learning) by Assessment Reform Group (2002), Lee and Coniam (2013) described the implementation FA in writing in terms of three phases: 1. Teachers' cooperative planning of the teaching resources and feedback forms; 2. instruction based on the teaching-learning cycle (setting the context, modeling and deconstruction of texts, joint construction, and independent construction) and 3. actual writing assessment phase using the same criteria at the instructional stage. In another study, Mak and Lee (2014) examined EFL teachers' implementation of FAoW in six classrooms over a course of one year through classroom observations and interviews with administrators and teachers. The schools adopted a FA plan with three phases of the writing process_ pre, during and post-writing stages. In pre-writing stage, teachers familiarized the students with the assessment criteria and set their goals. In the during-writing stage, the students benefitted from their peers' and the teacher's feedback and used their focused and coded corrective feedback. The feedback corresponded with the assessment criteria which had been established in the pre-writing stage. In the post-writing stage, the students recorded the number of errors in their error log and reflected on their progress. The act of reflection also involved students in thinking critically about their own writing and the feedback they received from both their peers and teachers so that they could make use of the information to feed forward and benefit their future writing. With the three staged research plan, the teachers were consequently able to teach what they assessed and assess what they taught.

In an earlier study by Tavakoli et al. (2018) based on Black and Wiliam's (2009) FA framework and Hattie and Timperley's model of feedback (2007), FAoW was operationalized in an instrument to measure EFL students' experience of FAoW practices as their role along with the teachers' in FA is of crucial significance and, according to many scholars (e.g. Feng, 2007), this role has been overlooked. Brookhart (2001) placed the students in the central role and, in line with Black and Wiliam (1998), considered assessment to be formative only when the information it provides is used for improving students' performance and learning. She explained that the limited research on the role of students in FA is probably because teachers are always considered to plan and administer classroom assessments. This study, as part of a bigger project, is to fill the gap and investigate the construct validity of a FAoW instrument which measures teachers' practice in the view of EFL learners in writing classes.

Assessment of writing has been documented by some researchers in the EFL context (e.g. Elahinia, 2004; Ghoorchaei, Tavakoli, & Nejad Ansari, 2010; Javaherbakhsh, 2010; Mosmery & Barzegar, 2015; Moradan & Hedayati, 2011; Naghdipour, 2016, 2017; Nezakatgoo, 2005; Sadeghi & Rahmati, 2017; Sharifi & Hassaskhah, 2011). Most of these studies on FA in writing classrooms have been experimental case studies on the effect of FA practices when introduced through an intervention or qualitative researches on the existing assessment practices in writing classrooms. For instance, Naghdipour's (2016) interviews with teachers and students showed that FA tools such as collaborative tasks, portfolio writing, and other process- and genre-based strategies were absent in the EFL writing classrooms. In another attempt (Naghdipour, 2017), FA was incorporated into a university EFL writing course and the data on students' beliefs and attitudes were collected through semi-structured interviews at the end of the semester and pre- and post-study attitude questionnaires (developed mainly in line with Lee, 2011). FAoW intervention was a three-session modular instruction to teach writing based on five FA strategies outlined by Black and Wiliam (2009). First, pre-writing stage of instructional tasks which made students write based on model essays, brainstorming and pooling of ideas, (see Naghdipour & Koç, 2015, for an overview). The second draft for each task was written in response to the peer-assessment and the third draft was revisions after the teacher assessment. FA intervention revealed an improvement in various aspects of participants' writing and development of their positive attitudes towards writing as well as FA.

There is a consensus in many studies on the beneficial effect that alternative forms of assessment. However, when implementing various forms of formative assessment is explored for writing classrooms, the existing researches fail to account theoretical models and operationalized set of FAoW practices for EFL context.

Operationalization of FAoW construct and the development of an instrument to measure FAoW was the focus of another study by Tavakoli et al. (2018). The information on the development of FAoW instrument is crucial to this research as this study aims at the construct validity of that instrument through a confirmatory approach and model fit. The instrument (Appendix I) was developed through an intuitive approach with 50 items in an earlier study. The items were classified under 5 factors (colored differently in Table 1) through a focus group interview with three EFL experts in the domain of assessing writing. The experts agreed on the five dimensions underlying the items in the instrument and indicated FAoW to be multidimensional. Table 1 illustrates the items under the five FA factors and the three writing stages.

Table 1. *FAoW framework, item and construct matching by experts (Adapted from Authors, 2018)*

	Where the learner is going? **Pre-writing (feed up)**	Where the learner is right now? **Writing (feedback)**	How to get there? **Post-writing (feed forward)**
Teacher	Items 1, 2, 3, 4, 5, 6, 7, 8, 12 **Clarifying criteria**	Items 14, 15, 18, 20, 22, 23, 29, 30, 31, 36, 40, 43, 48 **Evidence on students' current learning**	Items 32, 33, 37, 39, 41, 44, 45, 47, 50 **Feedback to move learners forward**
Peer	Items 9, 10 **Clarifying criteria**	Items 16, 17, 25,26, 28 **Peer-assessment**	
Learner	Items 11, 13 **Clarifying criteria**	Items 19, 21, 24, 27, 34, 35, 38, 42, 46, 49 **Autonomy**	

The notions of 'feed up, feedback and feed forward' corresponded with the main function of FA to "reduce discrepancies between current understandings and performance and a goal" (Hattie & Timperley's, 2007, p. 86). As Table 1 illustrates, thirteen items tapped pre-writing stage activities such as model-writing, pre-writing planning, setting writing goals, organizing and developing writing ideas, free writing, and clarifying assessment criteria. These writing activities related to 'feed up', defined as 'the goals one lays down to achieve' (Hattie & Timperley, 2007, p.86) and corresponded with clarifying criteria in Black and Wiliam's (2009) FA framework. Students set attainable goals so that they understand what they are working towards in the 'feed up' stage (i.e., where they are going).

'Feedback'/ while writing stage guided the second set of items in FAoW instrument and specified assessing the progress that was being made towards the goal. The items included writing practices such as process-writing/ multiple drafting, writing feedback on progress, peer-writing feedback, writing error log, computer feedback, autonomous writing revision, writing reflection and self-assessment. Thirty items were placed under this construct and tapped a variety of feedback (e.g. graded, focused, indirect, direct and descriptive) from various sources (e.g. peers, teachers and the learners). This stage of writing corresponded with 'where the learner is right now' principle of FA and implied the learners' prior progress and current state of learning.

Items for assessing students' performance at post writing stage encompassed those practices which could lead students for their future improvement such as reflection for future progress, teacher-oriented feedback and portfolio assessment. They corresponded with 'feed forward' in the feedback model and, as Mak and Lee (2014) confirmed, covered writing practices which gave students a direction of what they were to achieve in the future, a blueprint of where they were going in the future.

To date, factor structuring of FAoW based on a unified theoretical framework has received scant attention in the literature. This indicates the need to factor structure the construct of FAoW through CFA. Hence, this study aimed at both theory verification and modification in an EFL context and attempted to answer the following research questions:

1. Does the five-factor FAoW model fit the data collected from EFL students of writing courses?

2. What is a model describing EFL teachers' practice of FAoW in the view of EFL students?

2. METHOD

2.1. Participants

Since the researchers aimed at assessment of the students' writing assignments at discourse level, the criterion for selection of the participants for both research questions was their prior experience of writing tasks at the level of paragraphs and essays. This made the researchers select junior and senior university undergraduates and upper-intermediate or advanced level language school students in Iran. Based on the prior experience of writing assessment criterion, sampling of participants for both the interviews and the quantitative data collection was purposive.

For piloting FAoW instrument, three EFL students (two from language schools and one from a university) were selected. All the interviewees were female and the researchers resorted to the same criterion of having writing assessment experience in their selection.

For responding to both research questions, a purposive sample of 315 Iranian EFL students was selected from three non-state language schools and five universities. Of the initial cohort, 255 respondents had more reliably and completely filled in the instrument (response rate of 85%). The participants' age ranged from 13 to 48 (M= 22). Overall, sixty-seven of them were males and one hundred eighty-eight of the participants were female (See Table 2).

The selection of participants was based on the criterion of prior experience of writing assessment. Participants from universities were selected from senior and junior undergraduate students of English since the English curriculum in Iran requires students to pass three mandatory writing courses (Advanced Writing, Essay Writing and Paragraph Writing) in the first two years. All the participants had finished writing courses/lessons at discourse level and had experienced assessment of their essay writing tasks prior to completion of FAoW instrument. Furthermore, selection of students from language schools was based on the level of English textbooks they covered at the time of data collection (upperintermediate and advanced based on CEFR[†]) and the greement of researchers that the books and the school curriculums included writing tasks at the level of discourse.

As shown in Table 2, 66.2% of the participants were learning English writing through the university undergraduate curriculum (Teaching, Literature or Translation of English) and 33.7% of them through private language schools.

Table 2. *Participants Demographic Information by number (%)*

Context of education	N (%)	Gender		Education
		Male	female	
University*	169 (66.2)	28(16.6)	141 (83.4)	Teaching English= 47(27.8) English Literature= 27(16) English Translation=95(56.2)

Private Language School**	86 (33.7)	39 (45.3)	47 (54.7)	Highschool= 39(45.3)
				Diploma= 20 (23.2)
				Bachelor= 18(20.9)
				Masters=18 (20.9)

Total	255 (99.9)	67 (61.9)	188 (138.1)	

*Third and fourth year bachelor students of English

** Upperintermediate and advanced level students

[†] in CEFR (Common European Framework Reference), B2, C1 and C2 define upper-intermediate an advanced level of English

2.2. Instrument tool

FAoW instrument taps at students' experiences of FA practices in writing classrooms and their attitudes towards the helpfulness of each practice. It had initially been developed by Tavakoli et al. (2018) with five underlying constructs based on Black and Wiliam's (2009) Formative Assessment (FA) and Hattie and Timperley's (2007) feedback model. In line with the intuitive approach of scale construction (Hase & Goldberg, 1967), a comprehensive review of the literature was undertaken, and 50 Likert scale items were devised. Three experts in the field of writing assessment intuitively classified the items based on the five components of FA (clarifying criteria, evidence on students' learning, feedback to move learners forward, peer assessment and autonomy) and in three stages ("Where the learner is going/Pre-writing, "Where the learner is right now/Writing and "How to get there/ Post-writing"). The estimates for the current study were only derived from the students' responses to the four-point Likert scales under experience.

FAoW questionnaire which was developed by Tavakoli et al. (2018) and used in this survey consisted of two sections. Section I solicited details on the participants' demographics such as age, gender, writing and assessment experience and their highest level of academic qualification. Sections II was the items which sought to determine students' experience of and attitude towards FAoW. It was rated by EFL learners using a four-point Likert type scale for experience (ranged from 1 to 4) on the left and the scale of attitude (ranged from 1 to 5) on the left side of each item. In the experience scale, 1 was a practice that had never been experienced, 2 was rarely, 3 for often and 4 as a FAoW practice that had always been experienced by students in their writing classrooms. The attitude scale measured students' attitude from 1(very unhelpful) to 5 (very helpful).

While the development of the instrument and qualitative operationalization of its construct was the focus of the authors' earlier study (Tavakoli, et al., 2018), this study aimed at its quantitative construct validation. Here FAoW was piloted by three EFL learners before the large-scale administration and CFA in order to identify if the language of the instrument was comprehensible to EFL learners. Three EFL learners were interviewed separately; they were asked to read each item and explain or exemplify their understanding of each FAoW practice either in their first (Persian) or foreign language (English).

2.3. Data collection procedures

The development of the instrument and piloting it with the EFL learners took place in the first semester of 2015-2016 academic year. Afterwards, the interview with the three EFL learners were independently conducted. Each interview lasted 70 minutes on average, was audiotaped and transcribed verbatim for further analysis.

After the interviews, paper FAoW instrument (Appendix I) was distributed among EFL learners in both language schools and universities with the attendance of the first researcher to provide assistance in case required. There was no time restriction to complete the instruments, but filling out the instrument took approximately 35 to 45 minutes.

2.4. Data analysis

Using SPSS 19 Cronbach's index of internal consistency was estimated for internal consistency of FAoW instrument. To respond to the research questions, that is, to factor structure and verify Black and Wiliam's (2009) model and evaluate the model currently employed by teachers, CFA was run on the students' survey data using Analysis of a Moment Structures 22 (AMOS). To construct validate the instrument through CFA, the missing data was handled first. As 8.5% of the data (above 5%) were missing and MCAR test revealed nonrandom missing (Little's MCAR Test: Chi-Square= 3430.96, DF=3038, Sig. = .000), series mean method could not be used to handle missing data (Tabachnick & Fidell, 2007). Hence, single imputation had to be used to screen the missing data. The data were also checked for kurtosis, skewness, normality and outliers. Although the distribution of data was found to be normal for all variables with the skewness of all experiences within the acceptable range of +3 and -3, multivariate normality and linearity test revealed 28 outliers/cases (p<.05), which were removed from the subsequent analysis.

From several types of parameters which are commonly reported to indicate goodness of fit for measurement models, for evaluating the the FAoW model in this study we report one index for every of the three broad categories. Root Mean Square Error of Approximation (RMSEA) is reported for absolute fit which calculates the standardized residuals resulting from fitting FAoW model to the data. Comparative Fit Index (CFI) for relative fit is reported as it adjusts for the issues of sample size inherent in the chi-squared test of model fit and the normed fit index. CFI analyzes the model fit by examining the discrepancy between the data and the hypothesized model and indicates better fit when it is closer to 1. Finally, standardized root mean square residual (SRMR) was ultimately used as a parsimony fit index in this research. This is in line with Brown (2015) who advise researchers to consider and report at least one index from each category when evaluating the fit of their models. Because chi-square is the basis for most other fit indices, it is routinely reported in all researches as an original fit index for Absulute fit (Brown, 2015), it is reported in addition to the three indices.

While absolute fit indices do not use an alternative model as a base for comparison, relative fit indices compare a chi-square for the model tested (FAoW with five constructs) to “baseline”, “independence” model (Aka null model) with no latent variables in which all measured variables are uncorrelated. Parsimony-corrected fit indices are relative fit indices that are adjustments to most of the formerly-mentioned fit indices.

3. RESULT

3.1. The pilot study

Qualitative analysis of the interviews with the three EFL learners confirmed their understanding of the FAoW practices underlying the five constructs (Table 1) particularly with the help of definitions or examples which were provided for every item. In each interview, special attention was paid to the clarity of key terms which corresponded with the constructs. Although the language of the instrument was English, technical terms had been defined, exemplified or translated into the participants’ first language. Their verbal reports while reading each item and their admission at the end of interviews showed that despite the apparently confusing technical terms such as ‘assessment criteria’, outline or mind map’, ‘free-writing’, ‘descriptive feedback’, error log’, ‘portfolio’ and ‘qualitative feedback’, further definitions and exemplifications in the instrument extensively added to their understanding of the FAoW practices.

3.2. Descriptive statistics and the reliability of FAoW instrument

Cronbach’s index of internal consistencyrevealed an alpha value of .91 (Table 3), which suggests a high internal consistency for the instrument. In addition to the reliability for the sum scale, Cronbach Alphas were also computed for the five factors of FAoW, i.e. Clarifying criteria, Evidence on students’ current learning, Feedback to move learners forward, Peer-assessment and Autonomy, which, except for peer-assessment, showed an acceptable internal reliability (Values above.7 are considered acceptable, though values above .8 are preferable, Pallant, 2007) (Table 3).

Table 3. *Reliability and descriptive statistics for FAoW instrument (with 50 items in five factors)*

	Total Items	Five FAoW Factors				
		Clarifying criteria	Current learning evidence	Feed forward	Peer-assessment	Autonomy
Cronbach's Alpha	.91	.75	.72	.77	.60	.75
N of items	50	13	13	9	5	10
N of participants	255	255	255	255	255	255
Range	107.09	28.61	29	25	14	25
Minimum	70	18	17	10	5	10
Maximum	177	46.61	46	35	19	35
Mean	113	32.57	27.98	20.42	11.31	21.14
SD	19.48	6.01	5.28	4.82	3.10	5.13

Descriptive statistics in this table also shows the lowest mean (11.31) belonging to peer assessment; however, the decision was made to keep peer-assessment as an underlying section of the FAoW instrument for CFA analysis since, theoretically and based on Black and Wiliam (2009), it is considered as one of the sources of FA and a crucial agent among the three (teacher, peer, learner). Moreover, poor Cronbach alpha in peer assessment is statistically justified as it is attributed mainly to the few number of items (Pallant, 2007).

3.3. Confirmatory factor analysis procedures

Following Black and Wiliam's (2009) framework, a five-factor hypothetical model of FAoW was extracted. The number and nature of latent variables were based on its five components:

- clarifying criteria for success (feed up)
- eliciting evidence of students' current writing ability (feedback)
- providing feedback to move learners forward (feed forward)
- activating students as instructional resources (peer-assessment), and
- activating learners' as the owners of their own learning (Learner autonomy)

The five-factor recursive model was subjected to a confirmatory factor analysis to assess the goodness of model-data fit; the resulted model of FAoW with 50 observed variables/ items and five factors/ unobserved variables is illustrated in the following.

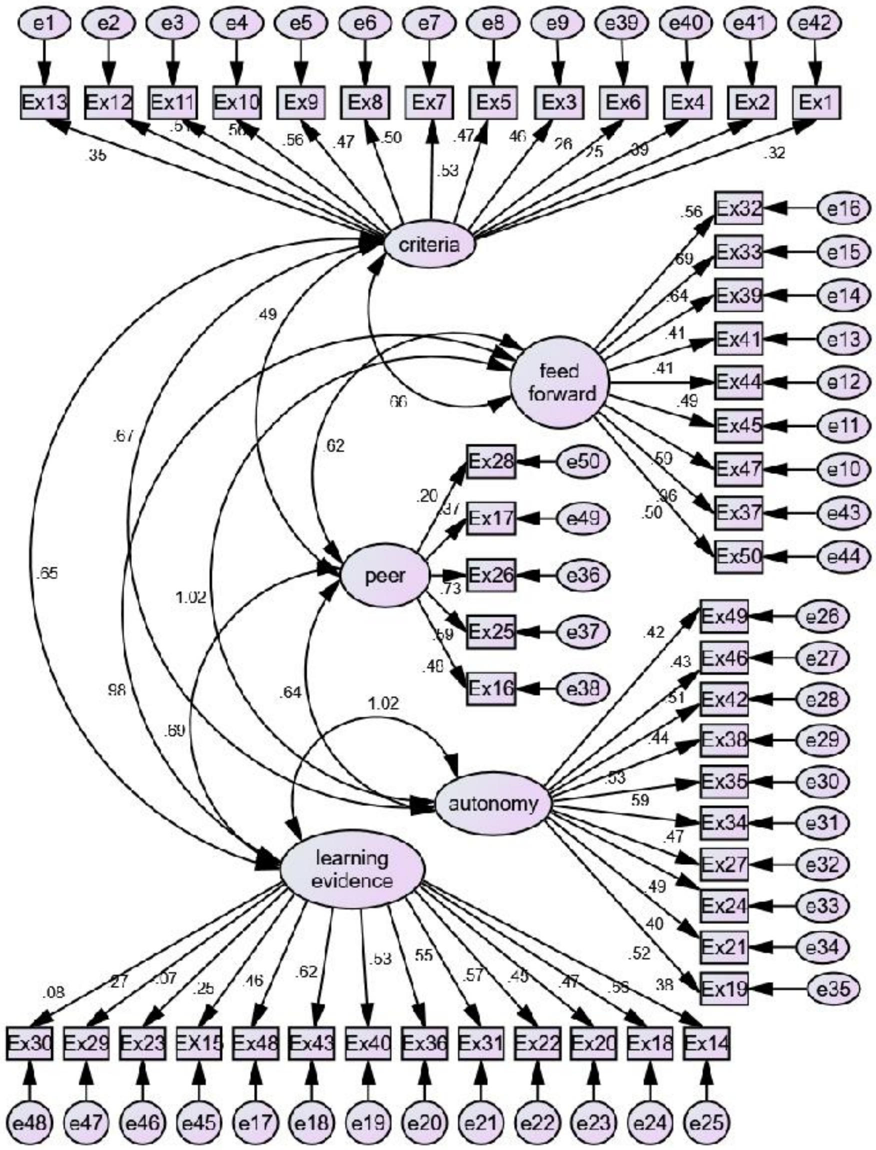

criteria
feed forward
peer
autonomy
learning evidence
Ex13 Ex12 Ex11 Ex10 Ex9 Ex8 Ex7 Ex5 Ex3 Ex6 Ex4 Ex2 Ex1
Ex32 Ex33 Ex39 Ex41 Ex44 Ex45 Ex47 Ex37 Ex50
Ex28 Ex17 Ex26 Ex25 Ex16
Ex49 Ex46 Ex42 Ex38 Ex35 Ex34 Ex27 Ex24 Ex21 Ex19
Ex30 Ex29 Ex23 EX15 Ex48 Ex43 Ex40 Ex36 Ex31 Ex22 Ex20 Ex18 Ex14

Figure 1. Five factor FAoW model

As Figure 1 shows, CFA model with five latent variables results in three main problems. Firstly, it showed high correlations between three latent variables, i.e. feed forward, autonomy and learning evidence/ feedback (r = 1.02, r = 1.2, r = .98), which is indicative of the three latent variables being only one factor rather than three. Graham, Harris, Fink and MacArthur (2001) explained low factor loadings between latent variables as indicator of a high discriminant validity. The issue of low discriminant validity was handled by merging feedback, feed forward and autonomy into one latent variable (with the label feedback) and trying a three-factor CFA model.

The second problem with five-factor model which encouraged the researchers to try three-factor solution was low factor loadings for eight items with a loading lower than .3. Items 4, 6, 15, 23, 28, 29, 30 and 37 respectively showed loadings of .25, .26, -.25, .07, .20, .27, .08 and .26. All the other loadings between the indicators and the latent factors as well as the covariance among the factors were significant at α= .001.

The third problem with this factor structure model was indices of fitness, particularly CFI, which was .70 and lower than the acceptable index (higher than .95, Hu & Bentler, 1999). As Table 4 illustrates, five-factor solution could show only acceptable CMIN, the root mean square error of approximation (RMSEA) and SRMR (respectively 1.85, .056 and .069). Small residuals (RMSEA_.06) indicate a small discrepancy between the observed correlation matrix and the correlation matrix estimated from the model (Hu and Bentler, 1999).

Five factor solution was not shown to have an acceptable comparative fit index nor discriminant validity. Hence, five factor solution of FAoW data statistically showed poor fit with the theoretical models of FA and writing feedback. The aforementioned problems with five-factor solution model made the researchers check three-factor solution through merging items under autonomy, feed forward and feedback and name them "feedback". Table 4 shows fitness indices and Figure 2 illustrates the model after modification.

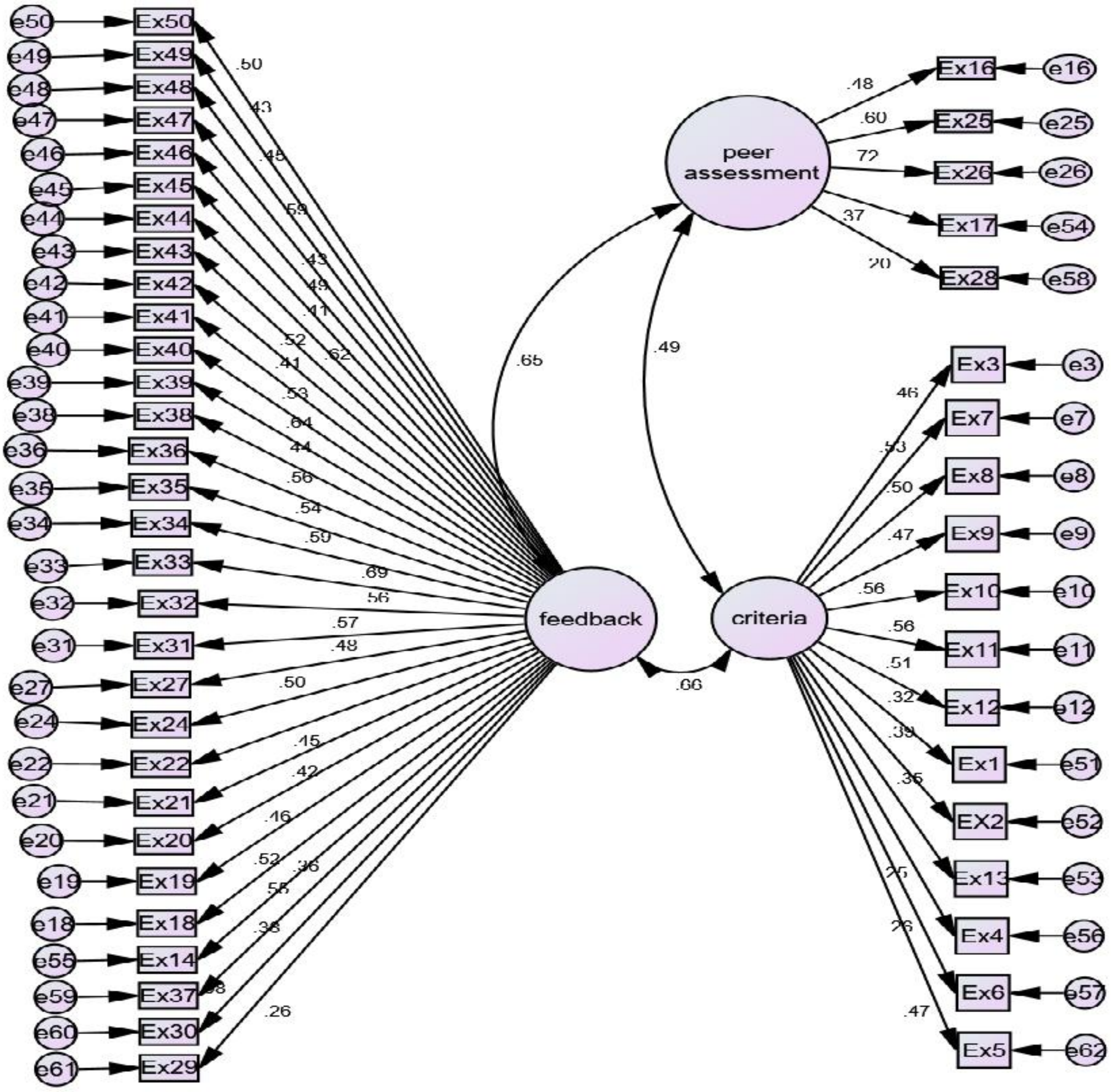


Figure 2. Three-factor model of FAoW

Table 4. *Fit indices for five and three factor CFA models of FAoW **

	Absolute fit Indices				Absolute fit Indices	Comparative Fit index	Parsimony Fit index
	Chi-Sq(x^2)	*Df*	*P* value	CMIN Chi-Sq(x^2)/D*f*	RMSEA	CFI	SRMR
Fitting Dataset for five-factor model	2161	1165	.00	1.85	.056	.70	.069
Fitting Dataset for three-factor model	1098	652	.00	1.62	.048	.84	.059
Acceptable threshold levels Hu and Bentler (1999)			P value>.05	1<CMIN<5	close to .06 or below	More than .95	Less than 0.08

*Modified FAoW after removals with factor loadings lower than .4

Although three-factor model resulted in higher discriminant validity with lower correlations latent variables/ factors (.66, .65 and .49), the problem of low standardized factor loadings remained in 8 items (4, 6, 15, 23, 28, 29, 30 and 37). In the modification process, the researchers maintained six items due to their relevance to the construct of FAoW and only removed two items (15 and 23) as they were reverse coded items introduced to the FAoW instrument to eliminate participants' guessing or boredom. More specifically, items 15 and 23 measured teachers' employment of one draft and product writing in contrast to process writing which taps FA. There is the argument in the literature (Brown, 2015) against using reverse coded items in questionnaires as they increase level of measurement error and affect loadings in factor analysis.

Comparison of model fitness indices between the five- and three-factor models of FAoW (Table 4) showed that the latter provided a better fit than the former; particularly in comparative fit index of CFI, which increased to .84, although not within the recommended acceptable range of above .95. The model has improved in discriminant validity as covariance between the three latent variables of feedback, feed up and peer assessment was relatively lower (.65, .69 and .49, Figure 2).

Hu and Bentler's (1999) evaluations criteria were employed for checking goodness fit between the target model and the observed data (see Table 4). Table 4 illustrates model fit indices for both three-factor and five-factor solutions. It reveals that probability or p-value is statistically significant and does not meet the acceptable range for model fit. With the sample size of more than 200, it is difficult to have a non-significant p value since x^2 statistic is very sensitive to sample size and is not relied upon as a basis for acceptance or rejection. Table 4 also shows that three of the indices (CMIN, RMSEA and SRMR) are within the acceptable range for model fit and confirm the absolute and parsimony fit of both models. Although CFI is lower than the acceptable value in both five and three-factor models (.70 and .84, respectively), three-factor model revealed a better fit in terms of CFI/ comparative fit index.

In response to the second research question, three-factor model can be considered as a more acceptable model in terms of goodness of fit for a better comparative index and higher discriminant validity, that is non- significant correlations between the latent variables (.62, .65 and .49 between peer assessment, feedback and criteria, see Figure 2). Three-factor model did not improve the low factor loadings with the aforementioned eight items either. Only one of the eight items improved in factor loading. Item 37 with a correlation coefficient of .26 under the latent variable of feedforward in five-factor solution gained a correlation of .38 under feedback in five-factor solution.

All in all, CFA revealed a poor fit for FAoW instrument which had been developed based on a Black and Wiliam's (2009) FA model with five factors and a writing model with three stages of pre-writing, writing and post-writing. Except for 13 items under prewriting (Where the learner is going) and the five items of peer assessment, all the other items under two stages of writing and post writing merged due to high correlation. In other words, items showing where the learner is right now functioned the same way that items showing how to get to the objectives.

3.4. The conceptual FAoW model

As the results of the research questions showed earlier, FAoW was not factor structured in the context of this study with the initially developed five constructs. The respondents' experience of FAoW supported three factors of setting criteria, feedback on students' writing tasks and peer-assessment. All the FAoW practices under feed forward and autonomy correlated statistically with the items which measured teachers' feedback on students' current learning (shown as learning evidence in Figure 1). This resulted in assessment in pre- and while-writing process, which is illustrated in a conceptual model in Figure 3.

The model encompasses three main stages of writing assessment and the FA practices that should be implemented on students' writing tasks: Prewriting FA practices, FA practices on the students' current writing task and post writing FA practices which can help students' future improvement and autonomy. As EFL learners' reports showed, the teachers explained learning goals and assessment criteria, encouraged them to brainstorm and develop an outline or mind map. These practices are all tightly related to FA and part of process writing (Hasim, 2014).

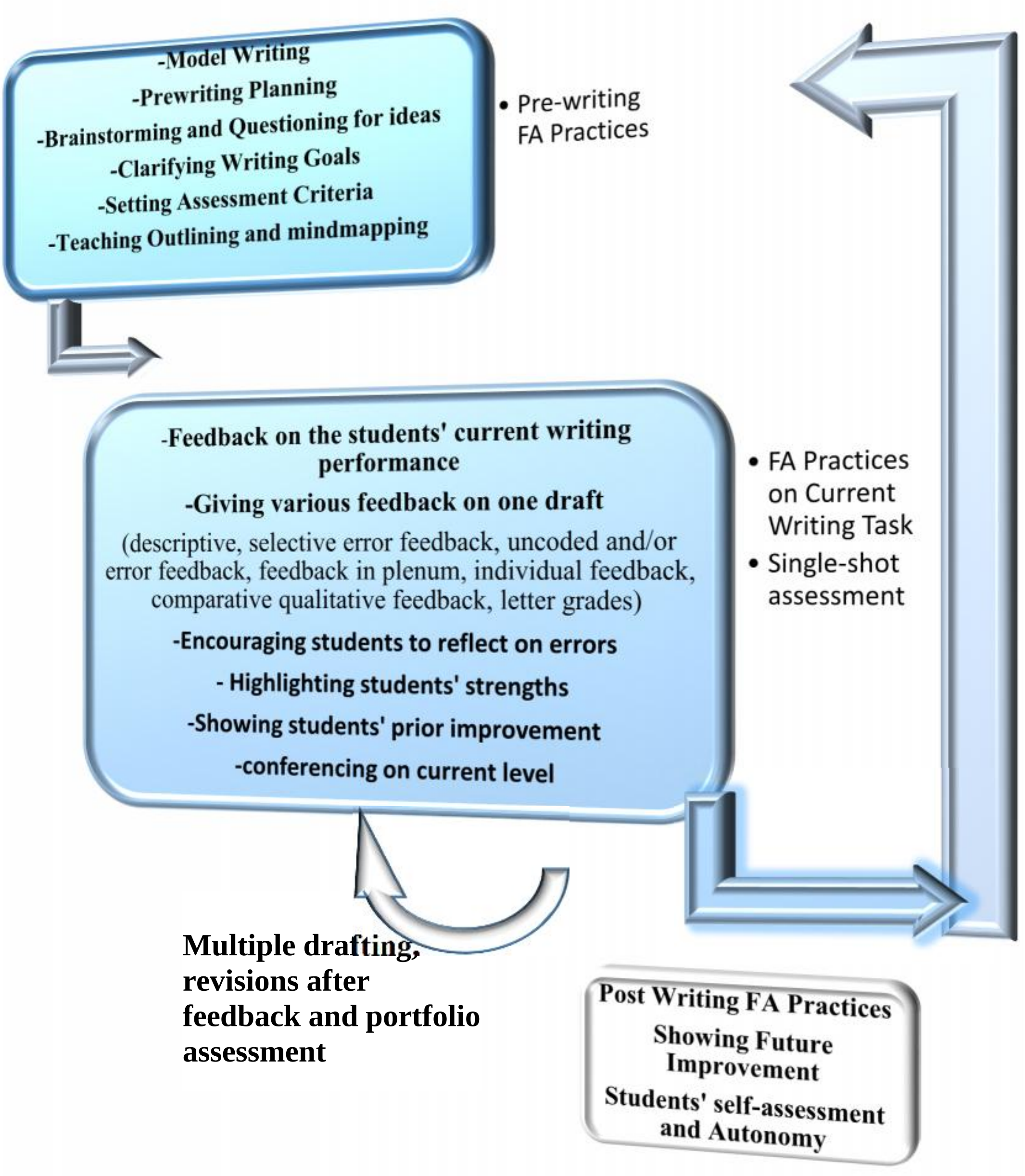

-Model Writing
-Prewriting Planning
-Brainstorming and Questioning for ideas
-Clarifying Writing Goals
-Setting Assessment Criteria
-Teaching Outlining and mindmapping
• Pre-writing FA Practices
-Feedback on the students' current writing performance
-Giving various feedback on one draft
(descriptive, selective error feedback, uncoded and/or error feedback, feedback in plenum, individual feedback, comparative qualitative feedback, letter grades)
-Encouraging students to reflect on errors
- Highlighting students' strengths
-Showing students' prior improvement
-conferencing on current level
• FA Practices on Current Writing Task
• Single-shot assessment
Multiple drafting, revisions after feedback and portfolio assessment
Post Writing FA Practices
Showing Future Improvement
Students' self-assessment and Autonomy

Figure 3. A conceptual model of teachers' practice of FAoW

The existing literature and the findings in this study, however, showed that the practices in the shaded gray parts of FAoW model in Figure 3 were implemented most frequently. In other words, the EFL students in this study did not think they achieved autonomy and independent self-assessment through post-writing FA practices. They learned about the writing goals in pre-writing stage and received single shot assessment on one draft rather than feedback on their revisions through multiple drafting. Hence, with the feedback, which had been usually direct error correction on a single draft, they moved to the next writing task in the next lesson. It seemed that they were deprived from the teachers' guidelines on how to improve and what to do next for the same task. Similar to studies in other EFL contexts (e.g. Saliu-Abdulahi, 2017), teachers delivered feedback to a finished text instead of asking for resubmission of the text for new assessment.

4. DISCUSSION and CONCLUSION

The specification of FAoW construct through models of FA and writing feedback and its operationalization was the initial stage of instrument development and the aim of an earlier study by the authors. In this study, the instrument was piloted through interview with three EFL learners for qualitative analysis of their comprehension. Subsequently, it was administered in large scale for factor structuring and construct validation through CFA.

The findings of the EFL learners' verbal report confirmed their understanding of the underlying constructs, with the help of examples, definitions and translation notes in the FAoW instrument. In line with Naghdipour (2017), Abdollahzadeh (2010) and Rahimi (2013), the interview findings showed that many of the FAoW practices had never been experienced by EFL learners and that product-based writing and teachers' direct error correction was very common among EFL teachers in writing classrooms. Abdollahzadeh's (2010) study did not aim at the construct of FAoW and only focused on writing strategies among the same population of undergraduate EFL students through large scale questionnaire survey and semi-structured interview. However, the metacognitive strategies in his study overlapped with many of the practices in pre-writing stage in FAoW instrument such as planning for writing, free-writing, awareness of writing purpose and brainstorming. The most common writing strategies among EFL learners were found by him to be metacognitive strategies, FAoW practices known as feed up in this study.

With reference to research questions in this study, our data could not fit in five-factor solution model and the construct of FAoW was found to have a better discriminant validity through three-factor solution. The three-factor model consisted of prewriting (setting assessment criteria) and writing (feedback and peer assessment) and left no post writing stage, which is equally, if not more, crucial in FAoW. The practices in prewriting stage formed criteria (known as feed up). Items under three factors (feedback on current writing, feed forward and autonomy) had to merge for a higher discriminant validity. In the literature, far too little attention has been paid so far to operationalizing the theoretical FA frameworks and writing models by accumulating a comprehensive list of formative feedback practices in writing. Carless (2007) similarly referred to this gap and the existing challenges in implementing the theoretical insights of FA from the literature.

Three-factor structuring was developed for two reasons, firstly due to the strong covariance of the items under feed forward, feedback on current state of writing and autonomy in five –factor structure and secondly because many of the items were theoretically measuring feedback while and after writing. The three highly correlated factors merged into one factor under the name of feedback as feedback was most inclusive of all the practices/ items. It seems that the student respondents in our study perceived the feedback they received on their writing tasks in writing stage as contributors to achieve autonomy and the ability to self-assess.

In addition to statistical justification, modifying the five-factor solution into three factor was theoretically plausible. The items under the three variables dominantly measured teachers' feedback in three stages, before, while and after writing and implementing them for achieving autonomy. Almost all of the items under the three merged factors were directly or indirectly measuring feedback. Furthermore, three stages of writing in Hattie and Timperley's (2007) model of feedback which had initially been used in the development of FAoW instrument could theoretically justify the possibility of merging three variables into one latent variable under the name of feedback and try FAoW model fit with three latent variables.

Abdollahzadeh's (2010) finding of higher frequency of metacognitive strategies (feed up in this study) can also corroborate our findings to the second research question as the practices or items under this construct were distinct from feedback and feed forward in writing and post writing stage. In other words, the students receive feedback on their writing performance in one stage; post-writing stage practices which move learners forward and make them more autonomous through reflection and self-assessment highly correlated with various forms of feedback which is given to students' current writing performance. This was also confirmed through the three participants in the interview. It probably indicated the misconception among EFL learners and maybe their teachers that single stage feedback can promote learners' writing ability to the level of autonomy.

A possible explanation for high correlation between 'current learning evidence' and 'autonomy' is EFL learners' experience of product writing which makes them believe they can progress and write more autonomously through various feedback that they receive in single writing drafts mostly in the form of direct error correction. It seems that their teachers set the criteria for assessment and showed the goals of writing in pre-writing stage; subsequent to the pre-writing stage they implemented all assessment feedback in one stage for students' single writing performance. Apparently, this way of assessment is believed to move learners forward and help them achieve independence and autonomy over time.

FAoW framework with five factors of FA and three stages of writing was not fit for the data collected in this study. Hence, it is probably possible to hypothesize that FA is not utilized in the current EFL context. This can partly be supported with the findings from Naghdipour (2016); Birjandi and Hadidi Tamjid (2012) and Rahimi (2009), who note that writing assessment in Iran follows a product-based tradition and feedback in a single stage. It is characterized by the teachers' focus on students' final products, which is not followed by students' further reflection on the received feedback. Many of the researches in Iranian undergraduate classrooms (e.g. Ketabi, 2015) and in other EFL contexts (e.g. Havnes, Smith, Dysthe & Ludvigsen, 2012; Saliu-Abdulahi, et al., 2017) confirm that assessment is not formative and lacks alternative approaches and various forms of FA.

Construct validation of FAoW instrument could not result in all the five underlying construct being confirmed by EFL learners in this study. Although the instrument was comprehensible for the participants in the qualitative phase of this study and seemed to be a valid measure for identifying students' experience of FAoW, when the factor structuring was analyzed for five factors, the model did not fit the data. For construct validity of FAoW instrument, three-factor solution could reveal a slightly better fit particularly in discriminant validity. The poor model fit of five-factor FAoW in the Iranian EFL context could suggest that the teachers set criteria and show objectives for the writing tasks in pre-writing stage, then incorporate feedback on the students' writing assignment, the feedback which is usually in the form of direct error correction of the form. Feedback is hardly utilized in this context to move learners forward. Feedback on one draft in the context of this study does not feature what Hawe and Parr (2013, p. 215) viewed as an effective practice to promote students' awareness about their improvement. Assessment in the context of this study is on "near-finished products" with the teacher fixing up mistakes not "developmental works in progress".

5. PEDAGOGICAL IMPLICATIONS

The findings of this study provide a set of FAoW practices suggesting an ideal FAoW model, which can be compared with what is actually being employed in EFL contexts. They complement the findings of earlier studies since they show that the practice of writing in EFL classrooms is single drafting and based on assessing the final writing draft. In addition to its theoretical contribution, this study has pedagogical implications for language education contexts. What seems to be missing in writing classrooms is showing the future trend and helping students how to revise the next drafts by implementing the feedback they have received. Traditional product-based approaches are still the frequent practice and the teachers often offer feedback on different aspects of the students' final draft at one time. The teachers need to encourage further drafting and revision of students' work.

FAoW instrument in this study was validated to identify the teachers' implementation of FAoW practices in the view of EFL learners; although the framework did not fit well with the data in this context, the instrument may have the potential to be utilized by other researchers in other contexts and writing classrooms as it is an operationalized model which can contribute to the utilization of FA. Hence it can be utilized by students, curriculum developers of writing programs and even the teachers (despite its wording) to evaluate the extent to which writing assessment is formative. If its construct is validated in other EFL and international contexts, the developed instrument can be used as a guideline for the teachers as well to know how FA is practiced. The instrument can also be employed by researchers as a classroom observation checklist to measure FAoW practices. The results of this study can additionally raise the awareness of those teachers who are not utilizing FA and are mainly concerned with showing learners their current state of learning rather than the future goals. The developed instrument can additionally pave the way for writing program designers and curriculum developers to implement FA in writing classrooms and utilize various assessment practices prior, while and after the writing stage.

FAoW is a vast area and can include any classroom activity as long as it aims to improve future performance. Multidimentionality of FAoW practices in the instrument was an inevitable problem for the researchers who aimed to develop an instrument with items which needed to tap a single dimension each. The researchers benefitted from both writing feedback model and FA as the theoretical foundation and sought to connect writing with FA. This could probably be assumed as one limitation in this study which could have affected the goodness of fit indices.

The generalizability of FAoW instrument as a measure to reflect teachers' practice of FA in writing classrooms is, therefore, subject to certain limitations. Poor goodness-of-fit statistics in this study makes generalizing the findings to the Iranian EFL teachers' very hard. Overall indices need to be locally justified through further research to provide more specific information about the acceptability and utility of the solution. These limitations made the researchers in this study consider caution when generalizing the findings and try to suffice to the conclusion that the assessment in writing classes in the context under this study seems to be practiced with three rather than five factors, clarifying assessment criteria and writing goals in prewriting stage, peer assessment and feedback in one stage to the final product.

Acknowledgement

This work was supported by NO funding agency.

ORCID

Elaheh Tavakoli https://orcid.org/0000-0002-3428-2603
Mohammad Reza Amirian https://orcid.org/0000-0002-3361-731X

6. REFERENCES

Arbuckle, J. L. (2012). Amos [Computer software]. Chicago, IL: SPSS.

Assessment Reform Group. (2002). *Assessment for learning: 10 principles. Research based principles to guide classroom practice*. London, UK: Retrieved from http://languagetesting.info/features/afl/4031afl principles.pdf

Abdollahzadeh (2010). Undergraduate Iranian EFL learners' use of writing strategies. *Writing & Pedagogy*, *2*(1), 65-90.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5-25.

Birjandi, P., & Hadidi Tamjid, N. (2012). The role of self-, peer and teacher assessment in promoting Iranian EFL learners' writing performance. *Assessment & Evaluation in Higher Education*, *37*(5), 513-533.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7-73.

Black, P. & Wiliam, D. (2006). *Assessment for learning in the classroom*. In J. Gardner (Ed.), Assessment and learning, 9-25. London: Sage.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability* (formerly: *Journal of Personnel Evaluation in Education*), *21*(1), 5.

Brown, T. (2015). *Confirmatory analysis for applied research* (2nd ed.). New York: The Guilford Press.

Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education: Principles, Policy & Practice*, *8*(2), 153-169.

Burner, T. (2015). Processes of change when using portfolios to enhance formative assessment of writing. *Assessment Matters*, *9*(2), 53-79.

Burner, T. (2016). Formative assessment of writing in English as a foreign language. *Scandinavian Journal of Educational Research*, *60*(6), 626-648.

Elahinia, H. (2004). *Assessment of writing through portfolios and achievement tests*. Unpublished Masters thesis, Teacher Training University, Iran.

Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. *Innovations in Education and Teaching International*, *44*(1), 57-66.

Feng, H. (2007). *Senior ESOL students' experiences of and attitudes towards formative assessment in mainstream secondary classrooms*. Unpublished Masters thesis, University of Canterbury, New Zealand.

Ghoorchaei, B. Tavakoli, M. & Nejad Ansari, D. (2010). The impact of portfolio assessment on Iranian EFL students‟‟ essay writing: A process-oriented approach. *GEMA Online Journal of Language Studies*, *10* (3), 35-51.

Graham, S., Harris, K.R., Fink, B., & MacArthur, C.A., (2001). Teacher Efficacy in Writing: A Construct Validation with Primary Grade Teachers, *Scientific Studies of Reading*, *5*(2), 177-202.

Hasim, Z. (2014). An integration of a process approach and formative assessment into the development of teaching and learning of ESL writing in a Malaysian university: A sociocultural perspective (Doctoral dissertation, University of Waikato).

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Havnes, A., Smith, K., Dysthe, O., & Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. *Studies in Educational Evaluation*, *38*(1), 21-27.

Hawe, E., & Parr, J. M. (2013). Assessment for Learning-Form and Substance in Writing Lessons. In *European Conference on Educational Research (ECER) Conference*. Istanbul, Turkey.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(*1*), 1-55.

Javaherbakhsh, M. R. (2010). The impact of self-assessment on Iranian EFL learners' writing skill. *English Language Teaching*, *3*(2), 213-218.

Johnson, R. C., & Riazi, A. M. (2017). Validation of a locally created and rated writing test used for placement in a higher education EFL program. *Assessing Writing*, *32*, 85-104.

Ketabi, S. (2015). Different methods of assessing writing among EFL teachers in Iran. *International Journal of Research Studies in Language Learning*, *5*(2), 3-15.

Lee, I. (2003). L2 writing teachers' perspectives, practices and problems regarding error feedback. *Assessing Writing*, *8*(3), 216-237.

Lee, I. (2007). Assessment for learning: integrating assessment, teaching, and learning in the ESL/EFL writing classroom. *The Canadian Modern Language Review, 64*(1), 199-213.

Lee, I. (2011). Formative Assessment in EFL Writing: An Exploratory Case Study, Changing English: *Studies in Culture and Education*, *18*(1), 99-111.

Lee, I. & Coniam, D. (2013). Introducing assessment for learning for EFL writing in an assessment of learning examination-driven system in Hong Kong. *Journal of Second Language Writing*, *22*(1), 34-50.

Mak, P., & Lee, I. (2014). Implementing assessment for learning in L2 writing: An activity theory perspective. *System*, *47*, 73-87.

Moradan, A., & Hedayati, N. (2011). The impact of portfolios and conferencing on Iranian EFL writing skill. *Journal of English Language Teaching and Learning*, *8*, 115-141.

Mosmery, P., & Barzegar, R. (2015). The effects of using peer, self, and teacher-assessment on Iranian EFL learners' writing ability at three levels of task complexity. *International Journal of Research Studies in Language Learning*, *4*(4), 15-27.

Naghdipour, B. (2016). English writing instruction in Iran: Implications for second language writing curriculum and pedagogy. *Second Language Writing Journal*, *32*, 81-87.

Naghdipour, B. (2017). Incorporating formative assessment in Iranian EFL writing: A case study. *The Curriculum Journal*, *28*(2), 283-299.

Naghdipour, B., & Koç, S. (2015). The evaluation of a teaching intervention in Iranian EFL writing. *The Asia-Pacific Education Researcher*, *24*(2), 389-398.

Nezakatgoo, B. (2005). *The effects of writing and portfolio on final examination scores and mastering mechanics of writing of EFL students*. Unpublished Master thesis, Allame Tabtba'i University, Tehran, Iran.

Pallant, J. (2007). SPSS Survival Manual, A Step by Step Guide to Data Analysis using SPSS for Windows, third edition, In *SPSS Survival Manual*. Open University Press, New York.

Pallant, J., & Manual, S. S. (2007). A step by step guide to data analysis using SPSS for windows. In *SPSS Survival Manual*. Open University Press, New York.

Rahimi, M. (2009). The role of teacher's corrective feedback in improving Iranian EFL learners' writing accuracy over time: Is learner's mother tongue relevant? *Reading and Writing*, *22*(*2*), 219-243.

Rahimi, M. (2013). Is training student reviewers worth its while? A study of how training influences the quality of students' feedback and writing. *Language Teaching Research*, *17*(1), 67-89.

Sadeghi, K., & Rahmati, T. (2017). Integrating assessment as, for, and of learning in a large-scale exam preparation course. *Assessing Writing*, *34*, 50-61.

Saliu-Abdulahi, D. (2017). Scaffolding writing development: How formative is the feedback?. *Moderna språk*, *111*(1), 127-155.

Saliu-Abdulahi, D., Hellekjær, G. O., & Hertzberg, F. (2017). Teachers'(Formative) Feedback Practices in EFL Writing Classes in Norway. *Journal of Response to Writing*, *3*(1), 31-55.

Sharifi, A., & Hassaskhah, J. (2011). The role of portfolios assessment and reflection on process writing. *Asian EFL Journal*, *13*(1), 192-229.

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

Wingate, U. (2010). The impact of formative feedback on the development of academic writing. *Assessment & Evaluation in Higher Education*, *35*(5), 519-533.

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 362–377
https://dx.doi.org/10.21449/ijate.592789

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

Mixture Rasch Model with Main and Interaction Effects of Covariates on Latent Class Membership

Tugba Karadavut [1,*], Allan S. Cohen [2], Seock-Ho Kim [2]

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

[2] The University of Georgia, Department of Educational Psychology (Quantitative Methodology), Athens, GA, USA

ARTICLE HISTORY

Received: 21 February 2019

Revised: 17 June 2019

Accepted: 02 July 2019

KEYWORDS

Covariates,
Item response theory,
Latent class membership,
Latent covariates,
Mixture Rasch model

Abstract: Covariates have been used in mixture IRT models to help explain why examinees are classed into different latent classes. Previous research has considered manifest variables as covariates in a mixture Rasch analysis for prediction of group membership. Latent covariates, however, are more likely to have higher correlations with the latent class variable. This study investigated effects of including latent variables as covariates in a mixture Rasch model, in presence of and in absence of interactions between the covariates. Results indicated the latent and manifest covariates influenced latent class membership but did not have much influence on class ability means or class proportions. The influence was relatively higher for latent covariates compared to manifest covariates. The effects of the covariates on class membership and on item parameters were class specific. Substantial effects of covariates on item parameters yielded smaller standard errors for item parameter estimates. A significant interaction term also had an effect on the coefficients for predicting and explaining latent class membership.

1. INTRODUCTION

A mixture Rasch model (MRM; Rost, 1990) assumes the examinee population is comprised of a finite number of discrete latent classes and a Rasch model with different item parameters possible within each class. The latent class portion of the model accounts for qualitative differences among examinees by detecting latent classes. The Rasch model part of the model accounts for quantitative differences among examinees both within and between latent classes. The MRM by itself detects the latent classes, but it does not explain why these classes form. This is necessary in order to understand why examinees are classified into different latent classes. Once latent classes are detected, therefore, a next step is to characterize each class to better understand the differences between classes. One method used for providing more information about these differences is addition of a covariate to the model in order to improve modeling of the association between the covariate and the latent class membership (Bilir, 2009; Cho, Cohen, & Kim, 2013; Choi, Alexeev, & Cohen, 2015; Dai, 2013; Smit, Kelderman, & van der Flier, 1999).

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

ISSN-e: 2148-7456 /© IJATE 2019

Different approaches can be adopted for inclusion of a covariate in a mixture model depending on the type (e.g., item specific covariates that refer to items) and level of the covariate (e.g., within level latent class covariates that are used to predict the latent class membership for a specific level in a multilevel model), or the parameter of interest to be predicted by the covariate (e.g., latent class membership, ability). The approach of extending mixture item response theory (IRT) models to include a multinomial logistic regression model with a covariate is adopted in this study to predict the latent class membership by using the covariate (Cho et al., 2013; Dai, 2013). The covariate in these models can be used as prior information (e.g., as an auxiliary variable) to predict the posterior probabilities of latent class membership.

Incorporation of a covariate in a mixture IRT model has been shown to be useful for detection of the latent classes and also for characterizing differences between the latent classes (e.g., Bilir, 2009; Choi et al., 2015; Dai, 2013; Smit et al., 1999). Previous research has included a single manifest categorical variable as a covariate in the model. Manifest covariates are not always sufficiently informative, however, they tend to be only moderately related to the variable causing the latent classes to form. In a differential item functioning (DIF) context, for example, manifest grouping variables were not very helpful for explaining causes of between group differences (Cohen & Bolt, 2005). In this study, we compare the effects of manifest and latent covariates with and without interactions on latent class membership. As an exploratory investigation, we tried to accomplish the purpose by presenting an application to data from the Program for International Assessment (PISA; OECD, 2013) mathematics literacy test.

In this paper, a finite mixture multinomial logistic regression structure with covariates was incorporated into a MRM for this purpose (cf. Cho et al., 2013). Latent covariates were expected to yield higher relationships with the latent class variable, because they were both obtained from examinee response data, albeit not from the same measures. This was expected to enhance the impact of the covariates on detection and subsequent characterization of the latent classes.

2. METHOD

2.1. Mixture Rasch Model (MRM) and Mixture Rasch Model with a Covariate (MRM-Cov)

Rost (1990) defined the probability of a correct response to item *i* by examinee *j* given that the examinee belongs to latent class g as:

$$P\left(X_{ii} = 1 \middle| \theta_j, g\right) = \left(\frac{\exp(\theta_{jg} - b_{ig})}{1 + \exp(\theta_{jg} - b_{ig})}\right), \qquad (1)$$

where θ_{jg} is the examinee's ability in class g, b_{ig} is the class specific item difficulty parameter, and X_{ij} is the observed response of examinee j to an item i. For model identification, $\sum_{i=1}^{I} b_{ig} = 0$ holds within each class. Bolt, Cohen, and Wollack (2002) noted that this norming constraint also makes θg comparable across classes and that the differences between the θ_g distributions can quantitatively explain the differences between the latent classes.

The MRM with a covariate (MRM-Cov) can incorporate a multinomial logistic regression model for π_{jg} in Equation 1 as follows:

$$\text{logit}\left(\pi_{jg}\right) = \beta_0 + \beta_1 y_j \qquad (2)$$

or similarly,

$$\pi_{jg} = \frac{\exp(\beta_0 + \beta_1 y_j)}{\sum_{g=1}^{G} \exp(\beta_0 + \beta_1 y_j)}, \qquad (3)$$

with the covariate y_j as the predictor, where β_{0g} is the intercept, and β_{1g} is the covariate effect in latent class g. The intercept and covariate effects in one of the latent classes were both fixed to zero for model identification (Cho et al., 2013).

2.2. Selection of Covariates

Smit et al. (1999) describe use of collateral information that has strong association with the latent class variable resulting in smaller standard errors on parameter estimates, when an equal or even fivefold smaller sample size was used in a MRM. Selection of a covariate that has strong association with the latent class membership, however, also requires theoretical as well as statistical justification. In most testing situations, manifest collateral information (e.g., demographic or contextual information) is available, since this type of information can easily be obtained through simple questionnaires or reference to institutional records.

Manifest variables, unfortunately, are not necessarily very useful predictors of latent class membership as the association between a manifest variable and the variable causing latent classes to form is typically modest at best (Cohen & Bolt, 2005). Further, the proportions of variance explained by these manifest variables are usually small even though they might be significant. Cohen and Bolt (2005) noted that latent variables, on the other hand, often have stronger relationships with the latent classes, thus providing more useful information regarding formation of the latent classes. Latent variables, however, typically require more complex substantive theories or statistical models for detection. In this study, we discuss using manifest and latent variables as covariates in the MRM analysis of PISA (OECD, 2013) mathematics literacy data. The student questionnaires from PISA provided collateral variables that were assessed for selection of latent as well as manifest covariates.

Strength of association between the covariates and the latent classes can be defined using bivariate probabilities of classification (Smit et al., 1999), or by using exponents of the coefficients (Dai, 2013). Two steps were used in the present study to determine appropriate latent covariates for incorporating in Equation 3: (1) the covariate selection, and (2) the MRM-Cov analysis with the selected covariates. In the covariate selection step, correlation coefficients were examined between candidate covariates and the latent variable of interest in order to determine the strength of association. In the second step, the exponents of the coefficients were determined as a measure of the association between the covariates and the latent classes in a MRM-Cov model. An empirical example is provided to demonstrate the two-step procedure for fitting a MRM-Cov model.

2.3. Empirical Example: Use of Latent Covariates to Predict Latent Class Membership in a Mixture Rasch Model

We illustrate this two-step procedure for selection and inclusion of covariates for predicting latent class membership in a MRM with two examples. The two examples included two different MRM-Cov models each including different combinations of latent and manifest covariates. The model in Study A included two covariates which did not have a significant interaction. The model in Study B included two covariates which did have a significant interaction. The purpose in these two studies was to gain insight about the effects of including more than one covariate in a MRM-Cov model on class membership in the presence of and in the absence of an interaction between the covariates. In addition, each study included a manifest and a latent covariate in order to compare their influence on the latent class membership.

2.3.1. *Data*

Data for the studies were taken from the 2012 edition of PISA (OECD, 2014) that assessed mathematics literacy as the main domain. Data from six English speaking countries (N=1,372) were used to mitigate differences due to translations (e.g., Bonnet, 2002): Australia (n_1 = 312), Canada (n_2 = 447), United Kingdom (n_3 = 289), Ireland (n_4 = 117), New Zealand (n_5 = 88), and the United States (n_6 = 119). PISA 2012 provided non-cognitive measures for students including manifest (e.g., demographic information, number of books at home) and latent variables (e.g., attitudes). Mathematics-related variables (e.g., attitudes towards math, beliefs about math) were considered for use as latent covariates. Booklet 5 was used for this example from the 13 booklets used for PISA 2012, because it included only mathematics items and most of its items required higher levels of cognitive process (e.g., employ, interpret) (OECD, 2014).

2.3.2. *Selection of Manifest Covariates*

A mathematics achievement score was calculated by summing the dichotomous item scores from the mathematics literacy test for each student. Among the manifest variables available with the PISA 2012 data, 31 were evaluated for possible use as covariates. These candidate variables were regressed on the raw mathematics achievement scores to find the most significant manifest variable that explained the largest proportion of variance. The purpose of this analysis was to find the manifest variable that was the best predictor of the mathematics achievement given the data. Regression analysis suggested that number of books at home was the best predictor of the mathematics achievement score (R-square = .124). Average hours a student spend each week on homework predicted the second highest proportion of variance explained (R-square change = .031).

2.3.3. *Selection of Latent Covariates*

PISA 2012 included four non-cognitive measures considered to be outcomes of mathematics education: (1) mathematics-related attitudes, beliefs and motivation; (2) general school-related attitudes and behaviors; (3) motivation to learn; and (4) educational expectations (OECD, 2013). Of these variables, those specifically dealing with mathematics-related attitudes, beliefs and motivation were considered as potential covariates. The mathematics-related attitudes included student interest in mathematics and student willingness to engage in mathematics. Student interest in mathematics included interest in mathematics at school, and intentions for further study in mathematics and in mathematics related careers. The willingness to be engaged was measured as "emotions of enjoyment, confidence and (lack of) mathematics anxiety, and the self-related beliefs of self-concept and self-efficacy" (OECD, 2013, p. 42). A mathematics-related attitude variable was considered as a potential covariate by combining the scales of the variables that comprised the mathematics-related attitudes. However, the scales of these variables were quite different for some of the variables, such as intentions and anxiety. As a result, two latent covariates were constructed: (1) self-related beliefs and (2) motivation. Items on these two latent covariates were scored on a four-point scale and were estimated using a partial credit IRT model (PCM; Masters, 1982).

2.3.4. *Student Motivation as a Latent Covariate*

PISA 2012 included scales measuring intrinsic and instrumental motivation, and short-term and long-term intentions to address the student motivation for mathematics (OECD, 2013). In this study, the eight-item intrinsic and instrumental motivation scale was used as an indicator of student motivation. The mathematics intentions measure was not included in the analyses since its scale did not combine meaningfully with the intrinsic and instrumental motivation scale. The coefficient alpha values for the four-item intrinsic motivation subscale and the four-item instrumental motivation subscale were both .90. The coefficient alpha for the eight-item student motivation scale was .92. Principal axis factoring yielded two factors that correlated .65. The two factors explained 62% and 15% of the variance, respectively. The factor loadings from an oblimin rotation with Kaiser normalization indicated that items on the instrumental motivation scale loaded on the first factor, and items on the intrinsic motivation scale loaded on the second factor.

2.3.5. *Self-Related Beliefs as a Latent Covariate*

Self-efficacy and self-concept are commonly used measures of self-beliefs in academic motivation research (Pajares & Schunk, 2001). Self-efficacy is described as a conviction or belief about one's ability to cope with certain tasks and self-concept is described as one's overall perception of his or her personal attributes evaluated by using continuous self-evaluation (OECD, 2013). A composite scale of self-beliefs was created by combining these two scales. The self-efficacy scale included eight items; the self-concept scale had five items. The coeffcieint alpha for the eight-item self-efficacy subscale was .86, and the coefficient alpha for the five-item self-concept subscale was .90. The coefficient alpha for the 13-item self-beliefs scale was .90. Principal axis factoring indicated two factors. The correlation between the two factors was .60. The factors explained 46% and 12% of the variance, respectively. Factor loadings from an oblimin rotation with Kaiser normalization indicated that the items of the self-efficacy scale loaded on the first factor, and items of the self-concept scale loaded on the second factor.

2.3.5. *Association between Covariates and Mathematics Achievement*

The association between the covariates and the mathematics achievement is shown in Table 1. The manifest variables of PISA 2012 (i.e., 31 manifest variables) together explained only 26% of the variability in the mathematics achievement scores. Self-beliefs, on the other hand, explained 29% of the variance in mathematics achievement by itself.

Table 1. Association between the Covariates and the Mathematics Achievement.

	Mathematics achievement	Index for the number of books at home	Motivation	Self-beliefs
Mathematics achievement	1.000	.359**	.253**	.535**
Index for the number of books at home		1.000	.065**	.158**
Motivation			1.000	.576**
Self-beliefs				1.000

** Correlation is significant at the 0.01 level (2-tailed).

Two linear regression analyses were done to predict mathematics achievement score. In Study A (i.e., covariates that did not have a significant interaction), the index of the number of books and self-beliefs were used as predictors. In Study B, motivation and self-beliefs were used as predictors. The regression for Study A did not yield a significant interaction between number of books and self-beliefs (β = -0.028, p = .639). The variables together explained 36% of the variance in mathematics achievement. In Study B, the covariates did have a significant interaction. The additive regression model, that is, the model with no interaction term for prediction of mathematics achievement using the self-belief and motivation scores, yielded a negative coefficient for motivation (β = -0.138, p < .001) and a positive coefficient for self-beliefs (β = 0.611, p < .001). Adding motivation to the model along with self-beliefs improved the relationship between self-beliefs and mathematics achievement but changed the sign of the coefficient for motivation indicating a suppression effect and, therefore, collinearity (Cohen, Cohen, West, & Aiken, 2003) between the variables. In this instance, motivation acted like a suppressor variable, as it had a weak positive correlation with mathematics achievement (r = .21) but a relatively strong correlation with self-beliefs (r = .57). In addition, as suggested by Cohen et al. (2003), this resulted as the correlation between mathematics achievement and motivation was less than the product of the correlation between mathematics achievement and self-beliefs, and the correlation between the motivation and the self-beliefs (e.g., .21 < .53 × .57 = .30).

Sequential regression analysis (also known as residual regression analysis) was used to account for the shared variance between the variables in the context of collinearity in the data. In this analysis, motivation was determined to be the important variable. Thus, self-beliefs was regressed against motivation. Self-beliefs was replaced with the residuals from this regression since the residuals represent the independent contribution of self-beliefs after accounting for motivation (Graham, 2003). In study B, therefore, motivation, the residuals that represented self-beliefs and the interaction of these two were used to predict mathematics achievement.

The initial analyses for covariate selection analyses showed that the effect sizes for motivation ($\beta = 0.197$, $p < .001$) and number of books at home ($\beta = 0.292$, $p < .001$) were smaller than that for self-beliefs ($\beta = 0.506$, $p < .001$). The interaction between the index of the number of books at home and self-beliefs was not significant ($\beta = -0.028$, $p = .639$). The interaction between motivation and self-beliefs was significant, although it had only a relatively small effect size ($\beta = 0.062$, $p = .026$) using Cohen's (1988) rules of thumb. Therefore, the anticipated effects of the motivation and number of books at home on latent class membership were also smaller relative to the self-beliefs. Small coefficients (e.g., approximately zero) from MRM-Cov were expected for the number of books at home and self-beliefs interaction given that it was insignificant.

2.3.6. *Estimation of the Model Parameters*

Estimation of the model parameters for each model was done using Markov chain Monte Carlo (MCMC) as implemented in the computer software OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009) (see Appendix A). The convergence of the model parameter estimates was assessed using three indices. Auto-correlations were examined as one indicator of MCMC convergence. In addition, the Monte Carlo error (MC error) for each posterior estimate was examined to determine if it was less than or equal to 5% of the standard deviation. Finally, the Heidelberger and Welch (1983) convergence diagnostic was used. Based on these indices, burn-in was determined to be 10% of the total of 30,000 iterations for each model except for the model with the number of books at home and the model with number of books at home and beliefs as covariates. History plots suggested a burn-in period of 5,000 and 4,000 for these latter two models, respectively.

2.3.6. *Estimation of the Self-Beliefs and Motivation Scale Scores*

The PCM was used for estimating self-beliefs and motivation. MC errors, Heidelberger and Welch (1983) and Geweke (1992) convergence diagnostics were used to inform convergence. The burn-in for the PCM was 30,000 iterations with a total of 150,000 post-burn-in iterations for estimation of self-beliefs. As some parameters had high autocorrelations, the chain was thinned to every 10th iteration resulting in 12,000 post-burn-in iterations used to obtain the posterior estimates. For motivation, the burn-in was 45,000 of a total of 225,000 iterations. The chain was thinned to every 10th iteration to reduce autocorrelations, resulting in 18,000 post-burn-in iterations.

3. RESULT / FINDINGS

3.1. MRM Analysis of the Mathematics Achievement Data

Schwarz's (1978) Bayesian information criterion (BIC) and Akaike's (1974) information criterion (AIC) were used to inform determining the number of latent classes in the models. BIC and AIC both suggested three latent classes for all models in both Study A (see Appendix B) and in Study B (see Appendix C).

To compare item parameter estimates between different models, mean and sigma equating (Marco, 1977) was used to transform the scale of models with a covariate to the scale of the model without a covariate. Additional transformation was not required for comparisons of the latent classes within the same model since the item parameters were mean centered within each class (e.g., Choi, 2014).

3.2. Results from Study A--Non-Interacting Covariates

In Study A (i.e., covariates that did not have a significant interaction), MRM model with the index of the number of books as a covariate (MRM-Cov-Books), MRM model with self-beliefs as a covariate (MRM-Cov-Self-Beliefs), and MRM models with the index of the number of books and self-beliefs as covariates with and without an interaction term (MRM-Cov-Self-Beliefs&Books) were estimated.

Squared errors within each class were calculated for comparing item parameter estimates from the model with a covariate (MRM-Cov) to the model without a covariate (MRM). In this study, the MRM was the baseline model. Squared errors were compared by taking square of differences between item parameter estimates from MRM model and item parameter estimates from MRM-Cov models for each latent class.

A factorial ANOVA was done to compare the log-transformed squared errors for item parameter estimates between the MRM-Cov models. The equal variances assumption was met using Levene's test ($F(11, 420) = 0.827$, $p = .613$). ANOVA results yielded a significant interaction between model type and latent class ($F(6, 420) = 5.924$, $p < .001$), with a small to moderate size eta-squared value of .065 based on Cohen's (1988) rules of thumb.

Pairwise comparisons of log-transformed squared errors for item parameter estimates between MRM-Cov models using Tukey's HSD procedure did not yield differences in mean squared error (MSE) values between the models for Class 1. This indicated the item parameter estimates from MRM-Cov models were similar to each other for Class 1. For Class 2 and Class 3, MSE values from MRM-Cov-Books were similar to the MSE values from MRM model, and smaller than the MSE values from the remaining MRM-Cov models. The MSE values from these remaining MRM-Cov models, on the other hand, were not different than each other. Similarly, the item parameter estimates from MRM-Cov-Books model were different than the item parameter estimates from the remaining MRM-Cov models, and the item parameter estimates from these remaining MRM-Cov models were similar to each other. This pattern was more evident in Class 3 than in Class 2. For all classes, the additive model without interaction and the model with interaction resulted in similar item parameter estimates.

A factorial ANOVA was conducted on the posterior standard deviations of item parameter estimates from different models to investigate whether a particular pattern existed for standard errors of item parameter estimates. A Box-Cox transformation ($\lambda = -0.656$) was applied to the standard deviations, as implemented in the R package MASS (Venables & Ripley, 2002). Levene's test ($F(14, 525) = 0.996$, $p = .456$) suggested equal variances. ANOVA results indicated a significant interaction between model type and latent class ($F(8, 525) = 7.695$, $p < .001$, $\eta^2 = .050$).

Pairwise comparisons of the posterior standard deviations of item parameter estimates were done using bootstrapping method with 10,000 samples because the cell means after Box-Cox transformation were not interpretable. For Class 1 and Class 2, the standard errors of item parameter estimates were similar from the different models. In Class 3, the standard errors of item parameter estimates were similar for the MRM and MRM-Cov-Books models. The standard errors of item parameter estimates were similar for the remaining models. The standard errors from the latter group of models were smaller than the standard errors from the former group of models.

The mean ability for Class 2 was fixed to zero for model identification, for each model. Class means for the different models were similar for Class 1. For Class 3, the class means from MRM and MRM-Cov-Books were similar to each other, and the class means from the rest of the models were similar to each other, although the differences in class means were trivial.

The mixing proportions did not exhibit a substantial covariate effect on the proportion of examinees in different classes for Class 1 as the mixing proportions from the different models were similar to each other. On the other hand, there was a clear pattern of effect for Class 2 and Class 3. This was similar to the effect observed for the mean ability estimates of Class 3. Specifically, the proportions of students in each class were similar for MRM and MRM-Cov-Books, and the proportions of students in each class were similar for the remaining three models. Incorporating self-beliefs in the MRM model as a covariate, or incorporating self-beliefs and the index of the number of books at home together with or without interaction resulted in an approximately 12% decrease in Class 2 membership and a 12% increase in the Class 3 membership. However, this did not result in a considerable change in membership to Class 1.

Coefficients from different models indicated that covariates did provide information for describing the latent classes (see Table 2). That is, the positive coefficients from MRM-Cov-Books showed that the students were more likely to belong to Class 1 and Class 3 as the number of books at home increased. The smaller coefficient for Class 1 indicated a smaller probability of being a member in this class as the number of books increase, compared to Class 3. Smaller coefficients also indicated that the number of books had a smaller effect size for predicting the class membership. The coefficients from the MRM-Cov-Self-Beliefs model also indicate that the students with higher self-related beliefs scores were less likely to belong to Class 1, and more likely to belong to Class 3. The exponents of the coefficients provide a measure of effect size in terms of odds ratios to indicate the effect of covariates on the latent class membership. The effect size for belonging to Class 3 (exp(2.908)) was higher relative to belonging to Class 1 (exp(-0.719)).

Inclusion of number of books at home and self-beliefs as the covariates in MRM without an interaction term yielded negative coefficients for both covariates for Class 1 and positive coefficients for Class 3. In other words, controlling for the number of books at home, the students with higher self-beliefs were less likely to be member of the Class 1 and more likely to belong to Class 3. Similarly, controlling for self-beliefs, students possessing higher number of books at home were less likely to belong Class 1 and more likely to belong Class 3. For Class 1, the coefficients for number of books at home and self-beliefs were similar to each other, which implies lacking of a differential covariate effect for this class. For Class 3, the effect size for self-beliefs controlling for number of books at home was 11.393 (=exp(2.958)/exp(0.525)) times the effect size for number of books at home controlling for self-beliefs. This exhibited a differential covariate effect for this class. Controlling for the effects of number of books did not cause a substantial change in coefficients of self-beliefs for both Class 1 and Class 3, compared to MRM-Cov-Self-Beliefs model. On the other hand, controlling for the self-beliefs caused a decrease in the coefficients of number of books at home for both Class 1 and Class3, compared to the MRM-Cov-Books model. This was consistent with results indicating smaller effect size for the number of books (β = 0.292) and a larger effect size for self-beliefs (β = 0.506) for predicting the mathematics achievement. Adding an interaction term to the MRM-Cov model with number of books at home and the self-beliefs did not result in a substantial change in the coefficients compared to the model without interaction. Further, the coefficient for the interaction term was approximately zero, consistent with the non-significant interaction term for predicting mathematics raw scores (β = -0.028, p = .639).

Table 2A. Coefficients from Different Models for Study A.

	MRM-Cov-Books		MRM-Cov-Self-Beliefs	
Class	Intercept	Books	Intercept	Beliefs
1	-0.288	0.904	-0.804	-0.719
2	0.000	0.000	0.000	0.000
3	-2.813	1.400	-3.804	2.908

Table 2B. Coefficients from Different Models for Study A.

	MRM-Cov-Self-Beliefs&Books (No Interaction)			MRM-Cov-Self-Beliefs&Books (Interaction)			
Class	Intercept	Books	Beliefs	Intercept	Books	Beliefs	Books*Beliefs
1	0.720	-0.793	-0.742	0.679	-0.791	-0.674	-0.041
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	-5.252	0.525	2.958	-5.574	0.631	3.231	-0.094

The MRM and MRM-Cov-Books had a high agreement of 93% for latent class assignment. The agreement between MRM and MRM-Cov-Self-Beliefs was 85%, and the agreement between MRM and MRM-Cov-Self-Beliefs&Books was 84% and 84% with and without interaction terms, respectively. The agreements in class membership for these pairs of models were similar to each other and smaller than the agreement between the classifications from the MRM and MRM-Cov-Books models. Overall, the results suggest that the covariates exhibited a considerable effect on class membership as the agreement between the MRM and MRM-Cov models changed considerably, depending on the covariate in the model.

3.3. Results from Study B--Interacting Latent Covariates

In Study B (i.e., covariates did have a significant interaction), MRM model with motivation as a covariate (MRM-Cov-Motivation), MRM model with self-beliefs as a covariate (MRM-Cov-Self-Beliefs), and MRM models with motivation and self-beliefs as covariates with and without an interaction term (MRM-Cov-Self-Beliefs&Motivation) were estimated.

Squared errors for item parameter estimates from the MRM model and the MRM-Cov models were calculated for each latent class as the square of the difference between item parameter estimates from the MRM model and from each of the MRM-Cov models. Factorial ANOVA analysis of natural log-transformed squared errors was conducted for comparing the item parameter estimates from MRM-Cov models. Homogeneity of variances assumption was met based on Levene's test (F(11, 420) = 1.042, p = .408). Results indicated a non-significant interaction between type of model and latent class (F(6, 420) = 0.810, p = .562). The main effects, however, were significant for both model type (F(3, 420) = 20.020, p < .001, η^2 = .107) and latent class (F(2, 420) = 38.270, p < .001, η^2 = .136), albeit with only moderate effect sizes.

Pairwise comparisons of MRM-Cov models using Tukey's HSD procedure did not yield significant differences in MSE values between the models for Class 1. For Class 2 and Class 3, MSE values between the MRM and the MRM-Cov-Motivation models were different than those between the MRM and the remaining MRM-Cov models. The MSE values from the remaining MRM-Cov models, however, were not different and were larger than those between the MRM and the MRM-Cov-Motivation model.

In other words, the item parameter estimates from MRM-Cov models were similar to each other for Class 1. For Class 2 and Class 3, however, the item parameter estimates from MRM-Cov-Motivation were more similar to those from the MRM model than they were to the remaining MRM-Cov models. Similarly, the item parameter estimates from these three remaining MRM-Cov models were similar to each other. For all classes, the additive model (i.e., without interaction) and the model with interaction resulted in similar item parameter estimates.

An ANOVA analysis was conducted on the posterior standard deviations of item parameter estimates from different models. A Box-Cox transformation (λ = -0.667) was applied to the standard deviations. Levene's test indicated that the equal variance assumption was met (F(14, 525) = 0.943, p = .511). Results indicated a significant interaction between model type and latent class (F(8, 525) = 6.224, p < .001, η^2 = .041). Pairwise comparisons between posterior standard deviations of item parameter estimates were done using bootstrapping with 10,000 samples. For Class 1 and Class 2, the standard errors of item parameter estimates were similar from the different models. In Class 3, the standard errors of item parameter estimates were similar for MRM and MRM-Cov-Motivation models. Likewise, the standard errors of item parameter estimates were similar for the three remaining models. The standard errors from the latter group of models were smaller than the standard errors from the former group of models.

The mean ability for Class 2 was fixed to zero for model identification, for each model. The class means from the different models were similar for Class 1. For Class 3, class means for ability appeared to be more alike for the MRM and MRM-Cov-Motivation models compared to class means for rest of the models, although the differences in class means were negligible.

The mixing proportions suggest that the proportion of students were similar across the models for Class 1. For Class 2 and Class 3, the mixing proportions were similar for MRM and MRM-Cov-Motivation and for the three remaining models. All three of these remaining models assigned more than half of the students to the second class. The inclusion of motivation as a covariate in the model classified roughly 4% of the students from Class 1 and Class 2 into Class 3 compared to the MRM. Incorporating self-beliefs in the MRM shifted approximately 12% of the students from Class 2 to Class 3. Adding self-beliefs and motivation together to the MRM model with or without an interaction term shifted about 12% of the students from Class 2 to Class 3.

Coefficients from different models exhibited a covariate effect for helping to characterize the latent classes (see Table 3). The positive coefficients from the MRM-Cov-Motivation model indicated that students were more likely to belong to Class 3 as the motivation score increases. The smaller coefficient for Class 1, on the other hand, indicated that the effect size for motivation was small. The coefficients in the MRM-Cov-Self-Beliefs model were negative for Class 1 and positive for Class 3 indicating examinees were less likely to belong to Class 1 and more likely to belong to Class 3 as their self-beliefs score increased. The model with both motivation and self-beliefs as covariates without an interaction term yielded a roughly zero coefficient for motivation and a negative coefficient for self-beliefs in Class 1. In other words, controlling for the motivation, the students with higher self-beliefs were less likely to be members of Class 1. On the other hand, controlling for self-beliefs, motivation did not show sufficient predictive power to estimate group membership for Class 1. The positive coefficients for Class 3 indicated that the students were more likely to be a member of Class 3 as either motivation or self-beliefs increased after controlling for the other variable. This tendency for self-beliefs was 8.551 (= exp(3.284)/exp(1.138)) times the tendency for motivation in odds ratio.

The model with motivation, self-beliefs and their interaction yielded coefficients different than zero for the interaction in Class 1 and Class 3. This was expected since the previous regression analysis yielded a significant interaction term between motivation and self-beliefs for predicting mathematics raw scores (β = 0.062, p = .026). For this model, controlling for self- beliefs and taking the interaction term into account, motivation did not show sufficient predictive power to estimate group membership for Class 1. Controlling for motivation and taking the interaction term into account, students with higher self-beliefs were less likely to be members of Class 1. Controlling for self- beliefs and taking the interaction term into account for Class 3, students with higher motivation scores were more likely to be members of this class. Similarly, students were more likely to be members of Class 3 as the self-beliefs increased. Controlling for the effects of motivation in the models with or without interaction, the coefficients for self-beliefs changed compared to the MRM-Cov-Self-Beliefs model in both Classes 1 and 3. Controlling for self-beliefs, on the other hand, did not cause a substantial change in the coefficients of motivation for Class 1 in the models with and without interactions compared to the MRM-Cov-Motivation. Controlling for self-beliefs, the coefficient for motivation differed somewhat for Class 3 both for the models with and without interactions. This suggested that using more than one covariate in the model helped explain class membership by taking into account the effect of the other covariate.

Table 3A. Coefficients from Different Models for Study B.

	MRM-Cov-Motivation		MRM-Cov-Self-Beliefs	
Class	Intercept	Motivation	Intercept	Beliefs
1	-1.446	0.084	-0.804	-0.719
2	0.000	0.000	0.000	0.000
3	-1.707	0.892	-3.804	2.908

Table 3B. Coefficients from Different Models for Study B.

	MRM-Cov-Self-Beliefs (Residualized) & Motivation (No Interaction)			MRM-Cov-Self-Beliefs (Residualized) &Motivation (Interaction)			
Class	Intercept	Motivation	Beliefs	Intercept	Motivation	Beliefs	Motivation*Beliefs
1	-1.642	0.049	-1.304	-1.357	0.052	-1.306	-0.879
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	-1.795	1.138	3.284	-1.826	1.140	3.277	-1.305

Agreement in class membership between the MRM and MRM-Cov-Motivation was as high as 94%. The agreement between the MRM and MRM-Cov-Self-Beliefs was 85% and between the MRM and MRM-Cov-Self-Beliefs&Motivation was 84% and 84% with and without interaction terms, respectively. The agreements in class membership for these pairs of models were similar to each other, and smaller than the agreement between the MRM and MRM-Cov-Books models. The patterns in the class membership agreement were similar to previous results. That is, the agreement between the MRM and MRM-Cov-Motivation models was greater than between the MRM and the remaining models. Incorporating self-beliefs and motivation in the MRM model together with or without interaction resulted in an approximately 13% decrease in Class 2 membership and a 13% increase in the Class 3 membership. However, this did not result in a considerable change in membership to Class 1. Results indicated a covariate effect on class membership causing students to shift between classes. This was likely because the agreement between the MRM and MRM-Cov models changed considerably depending on the covariate in the model.

4. DISCUSSION and CONCLUSION

This study was designed to investigate effects of use of a covariate in a mixture Rasch model (MRM) on latent class membership. In most testing situations, manifest variables such as demographic information can be obtained easily through short surveys following the administration of the test. The association between the manifest variables and the latent class variable, however, is generally moderate (e.g., Cohen & Bolt, 2005). Similarly, manifest variables in this study were found to account for a relatively small portion of the variance in mathematics achievement, even though they were significant predictors. Latent variables, on the other hand, explained a greater proportion of the variance showing the potential of the latent variables to be the better predictors of latent class membership compared to manifest variables, although some latent variables may be more useful than the others. Results of this study were consistent with previous research that latent covariates were more likely to have stronger associations with the dimension(s) along which the latent classes form. Contrary to the manifest variables, latent variables were also useful in this study for constructing meaningful composite scores based on previous research.

The results showed that the manifest and latent covariates did not have an impact on the number of underlying latent classes extracted, however, they helped explain the characteristics of the latent classes. The covariates changed the latent class membership proportions, however, they did not indicate a strong effect on class ability means. Latent covariates were more useful for explaining the characteristics of latent class membership compared to manifest covariates. Using more than one covariate did help explain the group membership after controlling for the other covariate. The effects of the covariates on latent class membership and on item parameters were class specific. Substantial effects of covariates on item parameters returned smaller standard errors for the item parameter estimates.

Results of this study suggested that incorporating more than one covariate in a mixture Rasch model should consider possible interactions between the covariates. Study A included covariates without a significant interaction, while Study B included covariates with significant interaction, although the interaction in Study B had a relatively small effect size. The models with interaction terms did not exhibit an effect on latent class membership proportions that was different from that for models without an interaction term. The significant interaction term in Study B, however, did show an effect on the coefficients for predicting and explaining latent class membership. It can be noted that the findings from this study are based on the two example studies that used empirical data and, hence, may not have direct applicability to other data or other sets of available manifested and latent covariates. More investigations, including simulation studies for which some parameters can be fully manipulated by researchers, are in need to check the generalizability of the findings.

ORCID

Tuğba Karadavut https://orcid.org/0000-0002-8738-7177
Allan S. Cohen https://orcid.org/0000-0002-8776-9378
Seock-Ho Kim https://orcid.org/0000-0002-2353-7826

5. REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723.

Bilir, M. K. (2009). *Mixture item response theory-MIMIC Model: Simultaneous estimation of differential item functioning for manifest groups and latent classes.* Unpublished doctoral dissertation. Florida State University.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement,* 39, 331-348.

Bonnet, G. (2002). Reflections in a Critical Eye: On the pitfalls of international assessment. *Assessment in Education: Principles, Policy & Practice, 9*, 387-399.

Cho, S. J., Cohen, A. S., & Kim, S.-H. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation, 83,* 278-306.

Choi, Y. J. (2014). *Metric identification in mixture IRT models.* Unpublished doctoral dissertation. University of Georgia.

Choi, Y. J., Alexeev, N., & Cohen, A. S. (2015). Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the TIMSS 2007 mathematics test. *International Journal of Testing*, *15*, 239-253.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*, 133-148.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York, NY: Routledge.

Dai, Y. (2013). A mixture Rasch model with a covariate: A simulation study via Bayesian Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *37*, 375-396.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & F. M. Smith (Eds.), *Bayesian statistics 4* (pp.169-193). New York, NY: Oxford Press.

Graham, M. H. (2003). Confronting multicollinearity in ecological multiple regression. *Ecology*, *84*, 2809-2815.

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*, 1109-1144.

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.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

OECD. (2013), *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris, France: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264190511-en

OECD. (2014). *PISA 2012 technical report*. Paris, France: OECD Publishing. Retrieved from www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Pajares, F., & Schunk, D. H. (2001). Self-beliefs and school success: Self-efficacy, self-concept, and and school achievement. In R. Riding & S. Rayner (Eds.), *Perception* (pp. 239-266). London, England: Ablex.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271-282.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Smit, A., Kelderman, H., & van der Flier, H. (1999). Collateral information and mixed Rasch models. *Methods of Psychological Research Online*, *4*, 1-13.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer.

Appendix A. OpenBUGS syntax for the MRM-Cov Models.

```
model
{
  for (j in 1:NE) {
   for (k in 1:NI) {
          r1[j,k]<-resp[j,k]
          r2[j,k]<-resp[j,k]
}}

#  1group model
for (j in 1:NE) {
      for (k in 1:NI) {
     tt1[j,k]<- exp(theta1[j] - beta1[k])
     p1[j,k]<-tt1[j,k]/(1 + tt1[j,k])
     r1[j,k]~dbern(p1[j,k])
     l1[j,k]<-log(p1[j,k])*r1[j,k]+log(1-p1[j,k])*(1-r1[j,k])
}
}
     loglik[1]<-sum(l1[1:NE,1:NI])
for(k in 1:NI){
b1[k]<-beta1[k]-mean(beta1[1:NI])
}

#  Priors for 1group
for (j in 1:NE) {
      theta1[j] ~ dnorm(0, 1)
}

for (k in 1:NI) {
    beta1[k]~dnorm(0,1)
 }

# 2group model
   for (j in 1:NE) {
       for (k in 1:NI) {
           tt2[j,k]<- exp(theta2[j] - beta2[gmem2[j],k])
           p2[j,k]<-tt2[j,k]/(1 + tt2[j,k])
           r2[j,k]~dbern(p2[j,k])
          l2[j,k]<-log(p2[j,k])*r2[j,k]+log(1-p2[j,k])*(1-r2[j,k])
           }
         gmem2[j] ~ dcat(pi2[j,1:G2])
         theta2[j] ~ dnorm(mut2[gmem2[j]],1)
        }
       loglik[2]<-sum(l2[1:NE,1:NI])

for (j in 1:G2) {
for(k in 1:NI){
b2[j,k]<-beta2[j,k]-mean(beta2[j,1:NI])
}}
# Priors for 2group
for (j in 1:G2){
    for (k in 1:NI){
          beta2[j,k]~dnorm(0,1)
         }
     }
for (j in 1:G2) {
      mut2[j]~ dnorm(0.,1.)
      }

#priors for coefficient
coef02[1] <- 0
coef12[1] <- 0

for ( i in 1:G2) {
    coef02[i] ~ dnorm(0,0.01)
    coef12[i] ~ dnorm(0,0.01)
}

for (j in 1:NE) {
    for (i in 1:G2){
      log(phi2[j,i]) <-  coef02[i]+ coef12[i]*books[j]
       pi2[j,i] <- phi2[j,i]/sum(phi2[j,1:G2])
    }
}

 for (i in 1:G2) {
for (j in 1:NE) {
      n2[j,i]<- equals(gmem2[j],i)
}
 sum2[i]<- sum(n2[1:NE,i])
  ppi2[i]<-  sum2[i]/NE
}


}
```

Appendix B1. Model Fit Indices for Study A.

	MRM		MRM-Cov-Books		MRM-Cov-Self-Beliefs	
Number of Classes	BIC	AIC	BIC	AIC	BIC	AIC
1	44510	44320	44510	44320	44510	44320
2	43690	43300	43730	43340	43800	43410
3	**43390**	**42800**	**43470**	**42870**	**43380**	**42790**
4	43640	42850	43720	42930	43650	42860

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; the smallest information criterion index is in bold.

Appendix B2. Model Fit Indices for Study A.

	MRM-Cov-Self-Beliefs&Books (No Interaction)		MRM-Cov-Self-Beliefs&Books (Interaction)	
Number of Classes	BIC	AIC	BIC	AIC
1	44510	44320	44510	44320
2	43800	43410	43810	43410
3	**43430**	**42840**	**43440**	**42850**
4	43710	42920	44310	43520

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; the smallest information criterion index is in bold.

Appendix C. Model Fit Indices for Study B.

	MRM-Cov-Motivation		MRM-Cov-Self-Beliefs(Residualized)-Motivation (No Interaction)		MRM-Cov-Self-Beliefs(Residualized)-Motivation (Interaction)	
Number of Classes	BIC	AIC	BIC	AIC	BIC	AIC
1	44510	44320	44510	44320	44510	44320
2	43730	43340	43810	43420	43810	43420
3	**43380**	**42780**	**43400**	**42810**	**43400**	**42810**
4	43630	42840	43660	42870	43680	42890

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; the smallest information criterion index is in bold.

International Journal of Assessment Tools in Education

2019, Vol. 6, No. 3, 378–395

https://dx.doi.org/10.21449/ijate.594749

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

Assessing the Relationship between Cognitive Load and the Usability of a Mobile Augmented Reality Tutorial System: A Study of Gender Effects

Emin Ibili [1,*], **Mark Billinghurst** [2]

[1] Department of Healthcare Management, Afyonkarahisar Health Sciences University, Afyonkarahisar, Turkey

[2] Department of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, Australia.

ARTICLE HISTORY

Received: 28.12.2018

Revised: 12.03.2019

Accepted: 30.06.2019

KEYWORDS

Augmented Reality,
Geometry,
Natural Interaction,
Usability,
Cognitive Load

Abstract: In this study, the relationship between the usability of a mobile Augmented Reality (AR) tutorial system and cognitive load was examined. In this context, the relationship between perceived usefulness, the perceived ease of use, and the perceived natural interaction factors and intrinsic, extraneous, germane cognitive load were investigated. In addition, the effect of gender on this relationship was investigated. The research results show that there was a strong relationship between the perceived ease of use and the extraneous load in males, and there was a strong relationship between the perceived usefulness and the intrinsic load in females. Both the perceived usefulness and the perceived ease of use had a strong relationship with the germane cognitive load. Moreover, the perceived natural interaction had a strong relationship with the perceived usefulness in females and the perceived ease of use in males. This research will provide significant clues to AR software developers and researchers to help reduce or control cognitive load in the development of AR-based instructional software.

1. INTRODUCTION

This paper explores the relationship between the usability of an Augmented Reality (AR) tutorial system (called ARGTS3D) and cognitive load. Cognitive Load Theory (CLT), has been frequently discussed in educational research over the last few decades and has undergone major developments over time (Klepsch, Schmitz and Seufert, 2007). One of the basic principles of CLT is to reveal the cognitive limitations that occur during information processing. According to Barrouillet et al. (2007), the cognitive load is the working memory load resulting from the amount of information that must be processed within a period of time. CLT is very important in the instructional design process because it exposes the structure of knowledge and the cognitive architecture in the process of this information. By evaluating learning environments

CONTACT: Emin İBİLİ ✉ eminibili@gmail.com Department of Healthcare Management, Afyonkarahisar Health Sciences University, Afyonkarahisar, TURKEY

ISSN-e: 2148-7456 /© IJATE 2019

instructional designers can reduce the cognitive load or manage the working memory load (Paas, Renkl & Sweller; 2003).

Three types of cognitive load is generally mentioned; intrinsic load, extraneous load and germane load (Sweller, Merrienboer & Paas, 1998). Intrinsic load corresponds to the cognitive load resulting from the usual complexity of the learning task, and it can be controlled by dividing the subject into smaller and simpler steps. For example, constructing toy blocks out of small cubes, so the assembly steps are divided into smaller and easier steps instead of giving the whole assembly step in a single scheme. Thus, instead of visualizing the whole process, starting by structuring small and easy tasks in the mind will contribute to the reduction of the inner cognitive load. When the preliminary information given to students is low, the number of elements to be processed in the working memory increases, which leads to an increase in the intrinsic load (Sweller, 2010). In this case, it is possible to reduce the intrinsic load by providing the necessary preliminary information for new learning (Klepsch, Schmitz & Seufert, 2007).

The extraneous cognitive load arises from the design of the learning material rather than the difficulty of the topic. Sometimes the instructional designer's use of teaching methods can make a subject more complex and provide distracting information. In this case mental resources are directed to unsuitable processes for the task and extracurricular cognitive load may increase (Kılıç, 2007). In contrast, many researchers have found that effective presentation methods based on CLT can make the learning processes more effective and efficient and thus reduce the cognitive load (Paas, Renkl & Sweller, 2003, Kılıç, 2007). For example, using written material to teach the motion of planets will make it difficult for the student to visualize the subject. However, using pictures will help students to more easily visualize the planet paths. Moreover, teaching videos and planetary movements will further contribute to understanding and visualization. In other words, the teaching method chosen by the teacher will encourage schema formation and facilitate the understanding of the subject or create more external cognitive load.

Germane load is based on mental information and diagrams that have been created for learning which the person has created based on previous experience. For example, a student who takes a foreign language for the first time will need more new schemes to construct the learning content in his or her mind. However, students with prior knowledge of learning content will build their learning on previous knowledge, thus reducing the formation of new schemes. Germane load is the working memory capacity that helps with conceptual learning by facilitating interaction with existing schemes associated with the intrinsic load (Sweller, 2010). Contrary to extraneous and intrinsic load, increasing the germane load is a desirable cognitive load type. This is because it facilitates cognitive load level by reducing intrinsic cognitive loading and facilitating the creation of correct mental diagrams (Paas & van Gog, 2006).

When users find it hard to understand multi-media educational systems, they can be distracted and this leads to them using different mental resources. In this case, the cognitive load increases and students can get confused using the system (Kılıç, 2007). There can be a lot of information and complexity which makes the user unsure where he or she is in the system and what they should do next (Kılıç, & Karadeniz, 2014). In order to increase the students' success in learning environments, they should be prevented from being overloaded and lost. For this purpose, it is useful to measure the cognitive load in order to determine whether the multi-media environments are effective and useful (Karadeniz, 2006).

In our research we are interested in how Augmented Reality (AR) can be used for teaching geometry, and how to do this in a way that minimizes cognitive load. AR is technology which seamlessly overlays virtual graphics on the real world in a way that both the real and virtual content can be interacted with at the same time (Kato & Billinghurst, 1999). AR applications typically use computer vision techniques to locate printed tracking markers onto which virtual objects are placed. AR has been shown to be effective for learning spatial information in a range of different domains, such as geometry (Cohen & Hegarty, 2014; Ibili & Sahin, 2015; Dünser, et al., 2006), anatomy (Jamali, Shiratuddin, Wong, & Oskam, 2015), health science (Moro, Štromberga, Raikos, & Stirling, 2017), tourism (Leue, Jung, & Dieck, 2015), retail (Poushneh, & Vasquez-Parraga (2017) and engineering (Wang et al., 2014), among others. However, more research needs to be conducted on the relationship between AR and cognitive load in a learning environment.

According to Bujak et al. (2013), cognitive activities which are not directly related to the learning objective create an additional cognitive load. This can be especially the case in AR applications which don't have an intuitive interface. For example, interacting with virtual objects using a mouse and keyboard in AR educational applications can create extra cognitive load and reduce learning effects (Bujak et al., 2013). However, AR interfaces can enable interaction with virtual content by using natural techniques that improve the usability of the system (Bujak et al., 2013, Wu, Hwang, Yang & Chen, 2018). One of these is virtual buttons, which enable touch based interaction with AR applications (Amaguaña, Collaguazo, Tituaña & Aguilar, 2018).

In this research, the following research questions were investigated:

- Does the relationship between perceived ease of use and the sub-factors of cognitive load differ according to gender?
- Does the relationship between perceived usefulness and the sub-factors of cognitive load differ according to gender?
- Does the relationship between the ease of use, perceived usability and perceived natural interaction differ according to gender?
- Does the relationship between the perceived natural interaction and sub-factors of cognitive load differ according to gender?

One of the main innovations of this research is to improve the usefulness and natural interaction level of virtual buttons with a matrix method. Using this method, teaching in the AR environment could be divided into smaller steps. According to Mayers 2005, small parts of instructional content can be used to reduce the internal cognitive load and allow the user to move between different content without being lost. The use of a large number of tracking markers in AR environments can create an extra cognitive load. Previous studies showed that the use of a complicated AR interface to interact with digital materials in the AR environment both creates an extra cognitive load in students and limits their natural interaction (Bujak et al., 2013; Wei, Weng, Liu and Wang, 2015; Lai, Chen and Lee, 2019; Ejaz, Ali, Ejaz and Siddiqui, 2019). For this reason, in this study the relationship between perceived usability and cognitive load factors was explored and the effect of the perceived natural interaction and gender in this relationship was investigated.

This research extends earlier work in cognitive load, Augmented Reality, and education. In this section we review this related work and discuss the research gap that our research addresses. Research on the effect of natural interaction interfaces on the usefulness of the system shows that both variables are strongly correlated. Kaushik and Jain (2014) emphasized that motion-based natural interaction interfaces will increase the perceived ease of use for the system. In addition, the researchers stated that this interface would provide the user with an interesting and remarkable user interface environment and provide more freedom to the user and increase the usefulness. Chessa and Noceti (2017), using AR scenarios, explored the naturalness of the movements of users in different environments. Researchers have found that manual interaction using Leap Motion gesture tracking in a stereoscopic environment is more similar to the interaction in the real-world scenario, and therefore this technique provides a high level of natural interaction. Xue, Sharma and Wild (2018) found that females with good computer knowledge who use Virtual Reality goggles and AR-based digital materials had a higher satisfaction score than males.

Extraneous load, also referred to as mental effort, occurs when the amount of unnecessary information in the learning memory increases and does not help learning. (Hsu, 2017). Intrinsic load refers to the natural complexity and difficulty of the learned content (Sweller and Chandler 1994). The germane load is related to the learning characteristics of the student and refers to the working memory resources so that the student can cope with the intrinsic load (Sweller, 2010). Costley and Lange (2017) stated that effective instructional design and presentation would contribute to the development of a high level of germane cognitive load and increased intention to use. They found that perception of ease of use was related to mental effort, and that a low level of mental load would positively affect behavioral intention by increasing the perceived usefulness and germane load. While the perceived usefulness of technology shows the perception of the student towards future performance, the perception of ease of use shows the intrinsic belief of the student's effort in using technology (Venkatesh & Davis, 2000).

Liou, Yang, Chen, & Tarng (2017) compared Virtusl Reality (VR) and AR-supported astronomy courses in terms of the cognitive skills and intentions of students. Researchers have stated that establishing a relationship between AR teaching materials and the real environment is easier than in the VR environment. The researchers reported that there is less cognitive load in AR environments and that AR environments directly contribute to the creation of cognitive schemas. They found that the benefits and attitudes perceived in AR environments were higher.

Arvanitis et al. (2011) stated that user comfort has an impact on the technology acceptance model factors, and that users' perception of limited motion when using the system has a negative impact on user satisfaction. Moreover, researchers have concluded that users spend less cognitive effort when they perceive the system as useful. Therefore, researchers stated that the development of natural interaction interfaces can positively affect the emotional, motivational and cognitive processes of the users. For example, Pantanoa, Rese and Baierc (2017) concluded that the use of AR-based mobile tourism is related to the perception of ease of use, and that the difficulty of the task negatively affects the effort and that the perception of ease of use positively affects the performance. Safadel (2016) found that the perceived interaction in AR environments was positively related to perceived usefulness and satisfaction.

Ismail et al. (2018) examined the effect of this teaching method on visualization and cognitive load levels of students by using an Augmented Reality supported instruction set. The researchers stated that AR-supported teaching increases the students' visualization skills and reduces their cognitive load levels. They also found that teachers were able to encourage students more easily and increase their motivation and academic achievement with an AR-supported teaching method. Lai, et al. (2019) designed an AR-based learning system to facilitate students' reading skills for science lessons. The researchers found that multimedia teaching significantly increased the learning achievement and motivation of primary school students; moreover, they found that extraneous cognitive load levels decreased significantly during learning activity. Fischini, Ababsa and Grasser (2018) have investigated the applicability of AR to aviation maintenance training tasks at various levels of expertise. The results show that the usefulness of AR was higher than the current system and had less cognitive load.

Ejaz et al. (2019) stated that if AR users make more cognitive efforts to use the system, they will be distracted and cannot focus sufficiently on the use of AR. Therefore, they stated that AR system design is important, especially for non-expert users. Khan, Johnston and Ophoff investigated the effects of AR technology on students' learning motivation. Researchers looked at the effect of AR mobile application on learning motivation using the ARCS motivating design model. They found that AR has a positive effect on motivation due to its interaction and multi-message design advantages.

Previous research has confirmed that gender is an important factor in the impact of technology on learning performance. For example, Lawton and Morrin (1999), found that males performed better in a simulated maze than females. Robertson (2012) stated that female students had better learning activity than males, because they spent more time writing dialogue in games they were playing. Similarly, there are conclusions that gender is effective in AR environments (Weiser, 2001, Sadi & Lee, 2015). Kimbrough, Guadagno, Muscanell and Dill (2013) stated that females are more interested in interaction with AR applications than males, and Cheng (2018) stated that gender can play a role in favor of female students in the relationship between scientific epistemic beliefs of students and their understanding of AR in the context of learning. Hsu (2017) has stated that male performance was higher than female because females learned to use AR later than males. Pantano, Rese and Baier (2017) found that the perception of ease of use of AR was equal in both genders, but females' satisfaction for AR use was higher than that of males. Ahmad and Goldiez (2005) concluded that males performed better in spatial visualization and orientation tasks than females.

Some previous studies have shown that the perception of natural interaction, which is considered one of the superior aspects of AR, is effective in reducing cognitive load (Bujak et al., 2013). The results of this research experimentally confirmed the assumptions about the effect of perception of natural interaction on cognitive load in AR environments. In addition, one of the important innovations of this research was to reveal the effect of system availability when designing AR teaching environments with natural interaction interface. For this reason, the relationship between the perception of natural interaction and the ease of use of the system was revealed in this study. Another important novelty of the study was to reveal the effect of gender in this relationship. Thus, instructional designers and researchers were given important clues while developing AR teaching environments with a personalized natural interaction interface.

2. METHOD

For our research we used ARGTS3D, AR geometry teaching software, developed by Ibili, Resnyansky and Billinghurst (2018). This is free software for Android mobile devices that can be downloaded from the Google Play store. ARGTS3D covers the 3D geometry topics taught in Turkey in the 8th grade, using approximately 70 AR teaching topics scripted, designed and developed by authors.

The subjects were divided into units and subheadings so that the students did not encounter excessive cognitive load while learning geometry subjects. In addition, appropriate interactive animations for each subject was created and virtual buttons were used for natural interaction with these animations. With the software, students had the opportunity to rotate, resize, zoom in, zoom out and move virtual objects.

The ARGTS3D software was developed with the Unity3D game engine and uses virtual buttons to support natural interaction. Virtual Buttons are areas in the real world that cause actions to happen when they are covered up by the user's hands. For example, if a user touches a virtual button then one of a virtual models in the AR scene might change shape or disappear. They often have a virtual image associated with them, looking like a real button, to show where the active area is. The Vuforia AR library provides support for virtual buttons, making it easy for developers to add this functionality to their projects (Amaguaña et al., 2018).

Figure 1, shows the home scene in the ARGTS3D AR application. One of the design intents of the ARGTS3D software was to create more natural interaction between the user and teaching materials. Kaptelinin and Nardi (2012) have stated that natural user interfaces should be easy to use, intuitive, fun, but not intrusive. For this reason, instead of using an AR tracking marker in this software, it aims to create natural interactions similar to the user's real environment interactions by using virtual buttons (the small model on the marker in Figure 1). Users can see the menu structure within the page in the blue menu in the right corner of the screen. When a virtual button is selected both the background color of the virtual button is changed and the representation icon's color in the right corner. By using virtual buttons on this page, the user can switch between six different unit scenes. Within each unit scene, there are virtual buttons that direct users to related subjects. By using the virtual buttons on the subject page, the content scene can be accessed where the AR materials of the related subject are displayed. After the AR material has been selected, it can be displayed on the left side of the main tracking marker, such as the yellow cube in Figure 1c. For a more complete description of ARGTS3D and how it is used see İbili et al. (2019).

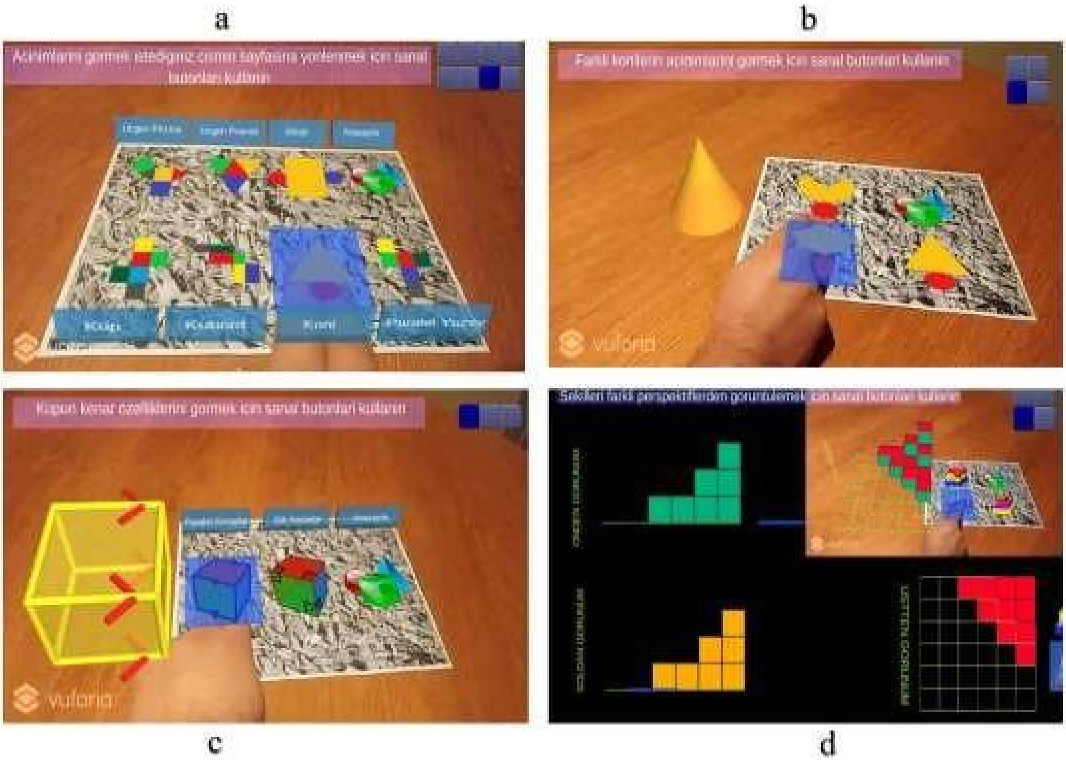


Figure 1. Augmented Reality sample scenes (İbili, Resnyansky and Billinghurst, 2019).

To test the ARGTS3D software, it was used by a teacher in two classes to teach secondary school three dimensional geometric topics to 59 students over four weeks. The demographics of the students included in the experimental study are presented in Table 1. Figure 2 shows students using the software in the classroom.

Table 1. Demographic profile results

Demographic Profile (N = 59, Age 13-14).	Category	Frequency	Percentage %
Gender			
	Male	29	49
	Female	30	51
Using ARGTS			
	with their own tablet or phone	43	73
	with their friends’ tablet or phone	16	27

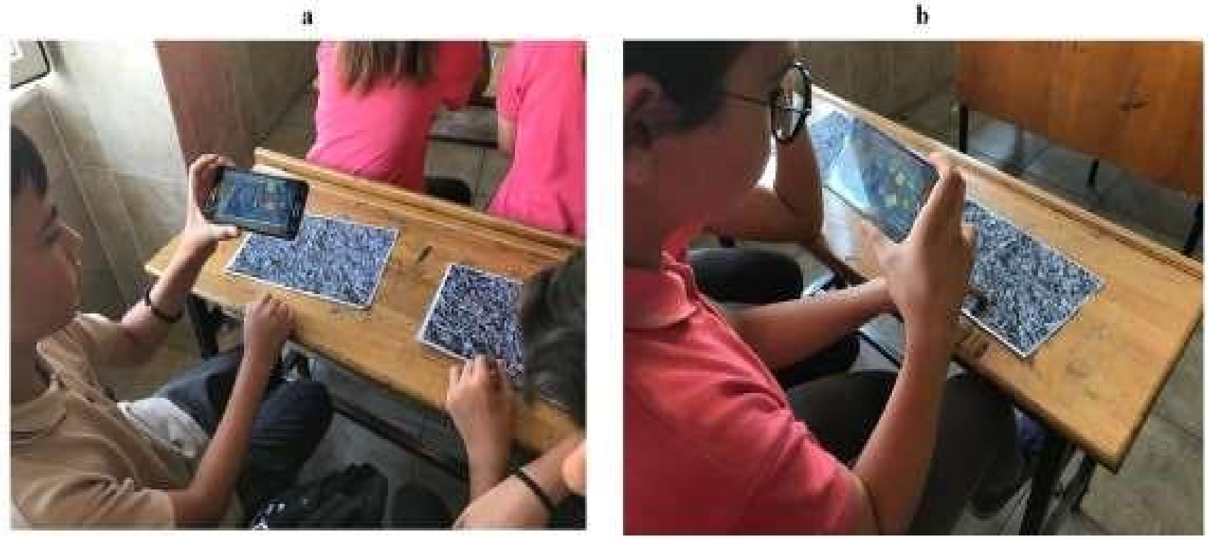


Figure 2. Example of experimental study process

In the teaching process, instruction was given for the following tasks;

- Students can draw nets of 3D objects and find out which prism shape a net belongs to.
- Students learn volume and area calculations by using unit cubes, start to establish connections between prisms, try to make structures with non-prismatic solids according to given volume, and predict the volume of rectangular prisms without using formulas.
- Students can draw two-dimensional views of three-dimensional objects from different sides, associate drawings made from different sides, and make isometric drawings.
- Students can recognize the pyramid, cylinder, cone, and vertical prism shapes and their structural elements.

In the experimental study, there was no intervention by the researchers about when the teacher should use the AR teaching software. We observed that the teacher usually used the software for about 15 minutes during the geometry lecture and question time.

2.1 Data Collection

To measure the cognitive load of the students as they used the software we used the cognitive load scale developed by Leppink et al. (2013). This consists of ten statements asked on a 10-point Likert-type scale between 1-strongly agree, and 10-strongly disagree (see Table 5 in the Appendix). The first three statements of the multidimensional cognitive load scale are about the intrinsic load, the next three items are related to the extraneous cognitive load and the last four items are developed for germane cognitive load measurement. In this study, the Cronbach Alpha reliability coefficients of the Cognitive Load Scale according to the dimensions were 0.70 for intrinsic load; 0.72 for extraneous load; 0.76 for germane load and 0.77 for the whole scale. This agrees with the results found by Leppink et al. (2013) who found Cronbach Alpha values of 0.81, 0.75, and 0.82 respectively.

A perceived usefulness and perceived ease of use questionnaire for collecting data was prepared using the Technology Acceptance Model (Davis, Bagozzi, 1989; Venkatesh, Davis, 2000; Agarwal & Karahanna, 2000). The surveys were first developed in English and later translated into Turkish (the students' mother tongue). A three-item questionnaire was prepared following the relevant literature review for the perceived natural interaction (see Table 5). However, the first item (NI1) in the natural interactıon factor was removed from the questionnaire after feedback from experts. The questionnaires used in the study were given to the students only at the end of the 4 weeks of instruction.

2.2 Data Analysis

The IBM SPSS 23 program was used for analysis using the arithmetic mean, standard deviation, an independent t-test and Pearson correlation coefficient. Before the analysis of the data and the interpretation of the findings, normality, linearity, and homogeneity assumptions were examined (Tabachnick & Fidell, 2001). The significance of the deviation of the distribution from the normal distribution for dependent variables was checked by using the Kolmogorow Smirnow test and the distribution was not deviated from the normal distribution (p> .05). The assumption of homogeneity of variance was also tested by using the Levene statistical test and it was found that the dependent variables of the study met the assumption of normality in each combination of independent variables (p> .05). Before the correlation analysis, the significance of the deviation of the binary scattering distributions from the linear distribution was calculated using the ANOVA coefficient and it was observed that the deviations of the paired correlations included in the analysis from the linear distribution were not significant and the analyses were continued with the parametric tests (p> .05).

3. RESULT / FINDINGS

Figure 1 shows the distribution of the mean scores of female and male students obtained from the cognitive load and usability scale sub-factors. As seen in Figure 1, the intrinsic load scores of both male and female students are below the average (5.5) and gender has no effect on the intrinsic load ($t_{(59)}$ = -.909, p> 0.05). This result shows that the intrinsic loads intended for 3D geometry subjects are manageable conditions at the end of the ARGTS3D supported geometry instruction. Similarly, the extraneous load scores of students are below average and gender has no effect on extraneous load ($t_{(59)}$ = -.830, p>0.05). This result shows that at the end of geometry teaching supported by ARGTS3D, there is a low amount of unnecessary knowledge in the learning memory of the student and this does not help learning. It also means that there is no effect of gender in the emergence of this extraneous load.

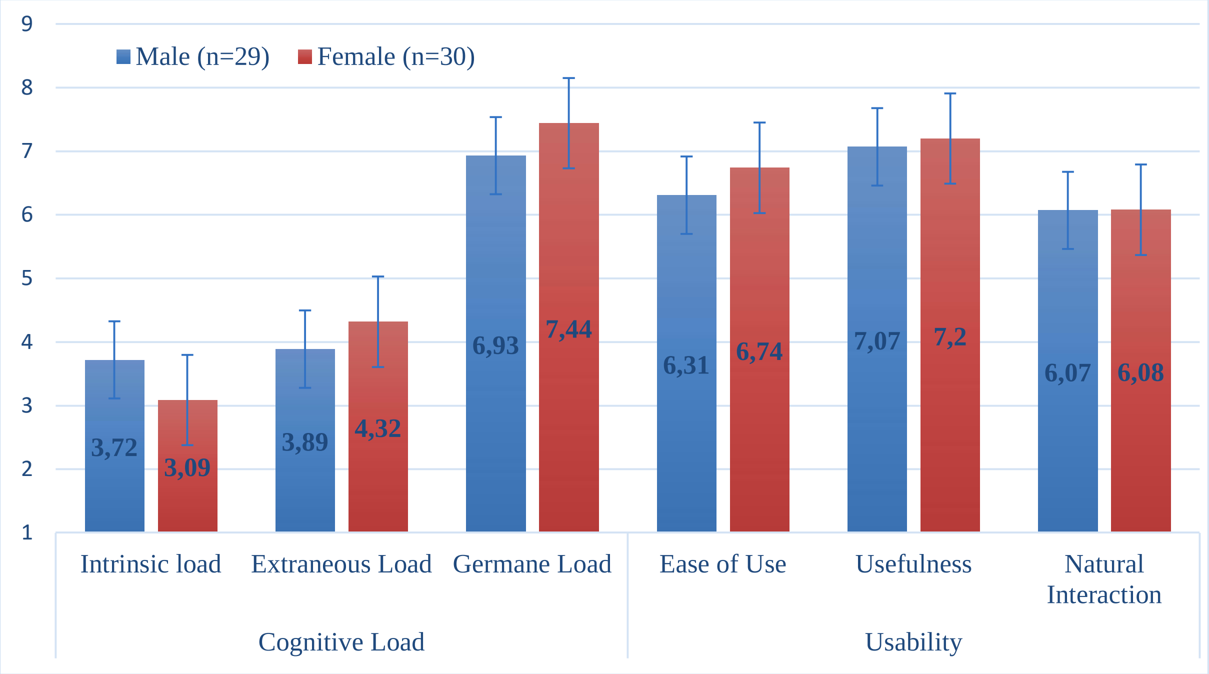


Figure 1. The average scores of the cognitive load and usability scores based on gender

On the other hand, contrary to the intrinsic load and extraneous load, the germane load is above average, and there is no significant difference according to gender ($t_{(59}$= -.797, p> 0.05). This result indicates that ARGTS3D assisted geometry teaching is effective on the germane load and increases the working memory resources used by the intrinsic load. In this way, it can be said that there is a decrease in both the internal load and the extraneous load, but the gender is not effective in increasing the germane load. The perceived ease of use ($t_{(59}$= -.667, p>0.05), perceived usefulness ($t_{(59)}$= -.241, p> 0.05) and perceived natural interaction scores ($t_{(59)}$= -.018, p>0.05) do not change according to gender and are above average. The results of the gender relationship between cognitive load scores and perceived ease of use and perceived usefulness are given in Table 2.

Table 2. The results of the correlation between cognitive load types and perceived ease of use and perceived usefulness

(N=59)	Usefulness		Ease of use	
	r	p	r	p
Intrinsic load				
Female	-.362	0.049*	-.155	.41
Male	-.213	.267	.046	.81
Extraneous load				
Female	-.119	.55	.009	.96
Male	-.371	.048*	.245	.20
Germane load				
Female	.782	.000**	.662	.000**
Male	.382	.037**	.646	.000**

**: 0.05 Significance level, **:0.01 Significance level.*

According to the results in Table 2, there was a negative correlation between the females' perceived usefulness and intrinsic load (r= -.362; p <0.05). On the other hand, there was a negative correlation between the males' perceived usefulness and extraneous load (r= .371; p <0.05). Also, the perceived usefulness was found to have a strong and positive relationship with the cognitive load for females (r= .782; p <0.01), and a moderate relationship for males (r = .382; p <0.05). In addition, it was found that the perceived ease of use had a strong and positive relationship with the germane load in both males and females (r_{female}:30=..662, r_{male}:29= .646, p< 0.01). The results of the relationship between perceived natural interaction with perceived ease of use and perceived usability are given in Table 3.

Table 3. The results of the correlation between the perceived natural interaction, perceived ease of use and the perceived usefulness.

	Natural Interaction			
	Female(n=30)		Male(n=29)	
	r	p	r	p
Ease of use	455	.015*	488	.007**
Usefulness	497	.005**	380	.048*

**: 0.05 Significance level, **:0.01 Significance level.*

According to the Table 3, there was a positive correlation between perceived ease of use and perceived natural interaction scores for both females (r = .455, p <0.05) and males (r = .488, p <0.01). In terms of usefulness scores, a strong positive correlation was found between the perceived usefulness scores and the perceived natural interaction scores for both females (r = .497, p <0.05) and males (r = .380, p <0.01). These results show that the relationship between females' natural interaction and ease of use is stronger, whereas the relationship between natural interaction and usability is stronger for males. The relation between perceived natural interaction and cognitive load subscale scores according to the gender are given in Table 4.

Table 4. Correlation between perceived natural interaction and cognitive load types

Cognitive Load	Natural Interaction			
	Female (n=30)		Male (n=29)	
	R	p	r	p
Intrinsic Load	-.447	.015*	-.173	.359
Extraneous Load	-.175	.3630	-.476	.008**
Germane Load	.639	.000**	.515	.004**

**: 0.05 Significance level, **:0.01 Significance level.*

According to the Table 4, there was a negative correlation between the intrinsic load and natural interaction (r = -.447, p <0.05) for males, whereas there was no significant relationship found for females (r=-.173, p> 0.05). Also, a strong negative correlation was found between the extraneous load levels and natural interaction (r = -.476, p <0.01) for females, but there was no relationship found between the extraneous load levels of the male students and natural interaction (r = -.175, p >0.05). In terms of germane load, a strong positive correlation was found between the germane load scores and the perceived natural interaction scores for both males and females (r_{female}:30=.515, r_{male}:29= .639, p< 0.01).

4. DISCUSSION

In this study, the relationship between the usefulness of AR teaching software and cognitive load was examined. In this context, the relationship between perceived usefulness, perceived ease of use and perceived natural interaction factors and intrinsic, extraneous, and the germane cognitive load were investigated. In addition, the effect of gender in this relationship was investigated and the following conclusions reached.

The intrinsic load scores and extraneous load scores of the 8th grade students for 3D geometry subjects were below the average. In addition, it was found that the germane load scores were above average and gender had no effect on cognitive load. It was seen that the complexity and difficulty perceived by the students in 3D geometry courses reached a manageable level at the end of the ARGTS3D supported geometry education. Euclidean geometry is insufficient to visualize 3D objects and students usually have difficulty in understanding and visualizing the concepts related to 3D geometry (Baki, Kösa, & Karakuş, 2008). According to Abdullah and Zakaria (2013), memorized geometry does not encourage students to think and remember.

Two-dimensional representations of knowledge therefore require more mental effort than three-dimensional representations (Wickens & Hollands, 2000). AR directs the working memory resources related to spatial visualization to the germane load using 3D representations, thus enabling information to be associated with each other and to relieve intrinsic load (Shelton, 2003, Nedim, 2013). It also has the potential to increase the germane load (Lee & Wong, 2014). Another contribution of AR in terms of cognitive load is that it keeps students active in the course because of allowing natural user interaction and thus contributes to reducing the extraneous cognitive load (Bujak, et al. 2013).

In some studies focusing on the effect of gender on perceived usefulness and perceived ease of use, different results were reported. For example, gender often had no effect on perceived usefulness and perceived ease of use, but according to some studies, females had a lower level of computer self-efficacy, so the females' perceived usefulness and perceived ease of use for new technology were adversely affected (Venkadesh and Morris, 2000; Ong ve Lai, 2006). In this study, it is assumed that self-efficacy perceptions are similar because all students have sufficient experience in using tablet, mobile phone and computers. Based on this assumption, it is thought that gender had no effect on perceived usefulness, ease of use and sub-cognitive factors for ARGTS use.

There was a negative correlation between the perceived usefulness of ARGTS3D and intrinsic load for females. A negative relationship was found between the perceived usefulness and the extraneous load for males. In addition, the perceived usefulness was strongly associated with germane cognitive load for females, whereas there was a moderate relationship between males. The perceived ease of use has a strong and positive relationship with the germane load for both male and female. Sweller (2010) emphasized that the intrinsic load is directly related to the working memory resources, and that as the extraneous load increases, the working memory load will increase and the intrinsic load will decrease. In addition, it is stated that the germane load increases the working memory resources used by the intrinsic load, so that the intrinsic load is reduced. According to Bhattacherjee (2001), the user's intention to continue to use the system depends on its expectation, satisfaction and perceived usefulness. Therefore, it can be said that when female students' perceptions about the usefulness of ARGTS3D increases, their intention to use the system will increase positively. As is seen in the results of the study, the females' germane cognitive load increased more than the males. Also, the extraneous cognitive load of males increased more than the females. Thus their usage of working memory resources decreased more. This has led to a reduction in the extraneous cognitive load in females and thus increased use of working memory resources. Therefore, the intrinsic cognitive load of the females was higher than males.

Ibili, Ryasnyansky and Billinghurst (2019) stated that ease of use and perceived usefulness are important determinants of satisfaction with AR learning system. They found that the effect of perceived usefulness on satisfaction is more effective than perceived ease of use. One of the most important reasons for this situation is that inexperienced users focus more on how to use the system, and experienced users focus more on the way they use the system (Xie, 2003). Based on these previous results (Xie, 2003; Ibili, Ryasnyansky and Billinghurst, 2019), it can be interpreted that female students' perceived usefulness of the ARGTS3D system decreases the intrinsic cognitive load by increasing use of satisfaction, frequency of use, and effective usage skills. Hou and Li (2014), in their research on the usefulness of educational mobile technologies, stated that male students focused on game-based attributes and female students focused on performance. Our results also supports Hou and Li’s research.

A positive relationship was found between the perceived ease of use and the perceived natural interaction scores for both males and females. However, this relationship was stronger for males than females. In contrast to this result, the relationship between perceived usability and perceived natural interaction was stronger for females than males. These findings are consistent with previous studies indicating that females mostly focus on the usefulness of the system and males focus on the usability of the system. (Xue, Sharma and Wild, 2018). In this context, it can be said that natural interaction affects satisfaction due to the perceived usefulness in females and decreases intrinsic cognitive load by increasing the performance-oriented usage frequency.

As the level of perceived natural interaction increased, the germane load of both female and male students was found to increase. This result could be caused by the quality of interaction and active learning environment that arises due to the user's experience with the ARGTS3D system. By using ARGTS3D, the student can rotate, resize, zoom in, zoom out and move virtual objects in the real environment. In this way, the student can both adapt the knowledge to his / her own cognitive structure and the attention of the student increases with the active learning environment (Bujak et al., 2013). According to Baraldi et al. (2009), using natural user interfaces in VR and AR applications can improve the quality of interaction (Baraldi et al.., 2009). Bujak et al. (2013) stated that interacting with AR-based virtual manipulatives led to further investigation of the learning content and encouraged students to learn. AR teaching environments allow for natural interactions, so the transparency of the interface between student and educational content increases (Bujak et al. 2013).

According to Leahy and Sweller (2005) if students use their imagination while learning concepts or procedures, the working memory resources are directed to related elements in the long-term memory, forming the core of knowledge, and the extraneous cognitive load is decreased. Lee and Wong (2014) emphasized that the ability to interact with teaching materials in AR and VR environments reduces the extraneous cognitive load by keeping students active and attracting their attention. Similarly, Bunch & Lloyd (2006) stated that the use of interactive maps reduced the extraneous cognitive load by attracting the student's attention. Klepsch, Schmitz and Seufert (2007) stated that cognitive load can be controlled by dividing the educational material into small and simple stages. In addition, the provision of preliminary information required for learning may reduce the intrinsic load. For this reason, the AR-assisted geometry education divided the subject into small steps with simple animations, effectively reducing the intrinsic load. In addition, the ability to easily access the information and materials needed by the students to recall and clarify their prior knowledge by means of virtual buttons has been effective in increasing the germane load and thus reducing the intrinsic load.

The research results show that as the level of perceived natural interaction increases, the intrinsic cognitive load in females and the extraneous cognitive load in males decreases. Cognitive load includes extraneous cognitive load and intrinsic cognitive load. Extraneous cognitive load occurs when the amount of unnecessary and unhelpful information in the student's learning memory increases (Paas, Van Gog & Sweller, 2010). The students' intrinsic cognitive loads increases when there is no relationship between the newly learned knowledge and the previous information. Therefore, the learning approach and the design of the instructional material affects the students' cognitive load (Young, Van Merrienboer, Durning & Ten Cate, 2014; Debue & Leemput, 2014). However, the effect of gender on intrinsic cognitive load and extraneous cognitive load may be caused by the ability of male and female to find information in memory through different methods.

Similarly, Bunch & Lloyd (2006) reported that males are more successful in activities such as mental rotation skills, and that females are more successful in tasks that require spatial information from long-term memory. Fabiyi (2017) and Gimba (2006) suggested that female students perform better in computation and spatial visualization than males. It can be said that female students use the resources of working memory more in order to respond to spatial tasks. It can also be said that when the level of natural interaction perceived in male students increases, students' attention and activity are increased, thus male students' extraneous load is decreased. However, the reduction of extraneous load in males had no direct effect on intrinsic load and indicates the presence of different variables that have an effect on intrinsic load. Sweller (2010) stated that the effect of extraneous load can be ignored in the case where the intrinsic load is manageable with working memory sources.

This research provides clues to AR software developers and researchers for reducing or controlling cognitive load in the development of AR-based instructional software. For example, the use of natural interaction interfaces has a positive effect on both the perceived usefulness of the AR teaching software and on the perceived ease of use. This means that the usefulness of AR teaching software can be increased by using natural interaction interfaces such as virtual buttons.

Natural interaction seems to be effective in reducing the extraneous load, and has a strong potential for students to keep active in the class and to focus on the lesson. In addition, natural interaction shows that interactions are effective in decreasing intrinsic cognitive load and increasing germane cognitive load. This result demonstrates the effect of AR-based instruction on both increasing and correlating the sources of working memory associated with intrinsic cognitive load. In addition, one of the most important results of this research is that gender effects the perception of usefulness and cognitive load in early school AR educational applications. In order to decrease the cognitive load in AR teaching environments, different AR teaching materials and techniques should be developed, taking into consideration the student's gender.

These research results reveal the importance of cognitive theory and multimedia design principles to be used when designing AR learning environments. However, the research has some limitations. This research data was limited to survey data obtained from 8th grade students after four weeks of using the ARGTS3D geometry education application. Therefore, the students' cognitive load factors related to the 3D geometry issues before and after using the application weren't compared. No information was collected about how often or for how long the users use this software in their extracurricular time. Therefore, individual differences in the effect of AR-supported instruction on cognitive load have been ignored. Other limitations include not comparing AR learning to non-AR learning, or exploring the use of other AR input methods, or the effect of using different AR displays such as head mounted displays (HMDs).

5. CONCLUSIONS

In this study, the relationship between the usability of the ARGTS3D application and cognitive load was examined. In this context, the intrinsic load, extraneous load and germane load were investigated for the ARGTS3D 3D geometry education tool. In addition, the relationship between these cognitive load factors and usability factors (perceived usefulness, perceived ease of use and perceived natural interaction) of the ARGTS3D software supported by virtual buttons were investigated.

One of the most important innovations of this research was the exploration of the effect of natural interaction perception on cognitive load and the effect on perceived usefulness and ease of use at the end of a four-week experimental process. For this purpose, a natural interaction factor was included in the study. Furthermore, in previous studies, theoretical research was conducted to determine the effect of the perception of natural interaction on cognitive load in AR environments. The results of this research is important because it validates these theoretical studies. Another innovation of this research is that the effect of perception of natural interaction on these variables differs according to gender.

The results of this study show that the perceived natural interaction has a strong relationship with perceived usefulness in female students and the perceived ease of use in male students. However, gender doesn't affect the perceived usefulness, perceived ease of use, and perceived natural interaction for the ARGTS3D teaching software. This result shows the presence of different variables other than natural interaction which effects the perceived usefulness and perceived ease of use. Similarly, there is a strong relationship between the perceived usefulness and extraneous load in men, while there is a strong relationship between the perceived usefulness and intrinsic load in women. In addition, both the perceived usefulness and perceived ease of use are strongly associated with germane cognitive load. However, the fact that the sub-factors of the cognitive load of the students do not differ according to gender indicates the existence of different variables that have an effect on these variables.

The focus of this study is to examine the relationship between the usability perceptions of AR teaching software and cognitive load factors in terms of gender. Therefore, the effect of ARGTS3D supported geometry teaching on cognitive load factors was not investigated. In the future we will design a research and control group to investigate the effect of AR supported teaching on cognitive load factors. In addition, we will explore the effect of different variables such as social norm, anxiety, self-efficacy and satisfaction on cognitive load using the

Technology Acceptance Model and supported with qualitative data. In addition to this, we will conduct new research to examine the effects of natural interaction and AR supported geometry with different display environments (such as head mounted displays (HMD), handheld displays (HDD) and desktop displays), and how this might affect student cognitive load levels. Also, the data related to usage frequency of the students should be collected from the system and the individual differences should be examined.

Acknowledgements

This research was supported by the postdoctoral research programme (BİDEB 2219) of The Scientific and Technological Research Council of Turkey (TUBITAK).

ORCID

Emin Ibili https://orcid.org/0000-0002-6186-3710
Mark Billinghurst https://orcid.org/0000-0003-4172-6759

6. REFERENCES

Abdullah, A. H., & Zakaria, E. (2013). Enhancing students' level of geometric thinking through van hiele's phase-based learning. *Indian Journal of Science and Technology*, *6*(5), 4432-4446.

Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly*, 665-694.

Ahmad, A. M., Goldiez, B. F., & Hancock, P. A. (2005, September). Gender differences in navigation and wayfinding using mobile augmented reality. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(21), 1868-1872). Sage CA: Los Angeles, CA: SAGE Publications.

Amaguaña, F., Collaguazo, B., Tituaña, J., & Aguilar, W. G. (2018, June). Simulation System Based on Augmented Reality for Optimization of Training Tactics on Military Operations. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*, (pp. 394-403). Springer, Cham.

Arvanitis, T. N., Williams, D. D., Knight, J. F., Baber, C., Gargalakos, M., Sotiriou, S., & Bogner, F. X. (2011). A human factors study of technology acceptance of a prototype mobile augmented reality system for science education. *Advanced Science Letters*, *4*(11-12), 3342-3352.

Baki, A., Kösa, T., & Karakuş, F., Çakıroğlu, Ü (2008). Uzay geometri öğretiminde 3D dinamik geometri yazılımı kullanımı: öğretmen görüşleri. *In International Educational Technology Conference, Eskisehir, Turkey* (pp. 6-9), 2008, May.

Baraldi, S., Del Bimbo, A., Landucci, L., & Torpei, N. (2009). Natural interaction. In *Encyclopedia of Database Systems* (pp. 1880-1885). Springer, Boston, MA.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 570.

Bhattacherjee, A. (2001). Understanding information systems continuance: an expectation-confirmation model. *MIS quarterly*, 351-370.

Bujak, K. R., Radu, I., Catrambone, R., Macintyre, B., Zheng, R., & Golubski, G. (2013). A psychological perspective on augmented reality in the mathematics classroom. *Computers & Education*, *68*, 536-544.

Bunch, R. L., & Lloyd, R. E. (2006). The cognitive load of geographic information. *The Professional Geographer*, *58*(2), 209-220.

Cheng, K. H. (2018). Surveying Students' Conceptions of Learning Science by Augmented Reality and their Scientific Epistemic Beliefs. *Eurasia Journal of Mathematics, Science and Technology Education*, *14*(4), 1147-1159.

Chessa, M., & Noceti, N. (2017). Investigating Natural Interaction in Augmented Reality Environments using Motion Qualities. In *VISIGRAPP (6: VISAPP)* (pp. 110-117).

Cohen, C. A., & Hegarty, M. (2014). Visualizing cross sections: Training spatial thinking using interactive animations and virtual objects. *Learning and Individual Differences*, *33*, 63-71.

Costley, J., & Lange, C. H. (2017). Video lectures in e-learning: effects of viewership and media diversity on learning, satisfaction, engagement, interest, and future behavioral intention. *Interactive Technology and Smart Education*, *14*(1), 14-30.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management science*, *35*(8), 982-1003.

Debue, N., & Van De Leemput, C. (2014). What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in psychology*, *5*, 1099.

Dünser, A., Steinbügl, K., Kaufmann, H., & Glück, J. (2006, July). Virtual and augmented reality as spatial ability training tools. In *Proceedings of the 7th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction: design centered HCI* (pp. 125-132). ACM.

Ejaz, A., Ali, S.A., Ejaz, M.Y & Siddiqui, F.A. (2019). "Graphic User Interface Design Principles for Designing Augmented Reality Applications" International Journal of Advanced Computer Science and Applications(IJACSA), 10(2), 209- 216, http://dx.doi.org/10.14569/IJACSA.2019.0100228

Fabiyi, T. R. (2017). Geometry concepts in mathematics perceived difficult to learn by senior secondary school students in Ekiti State Nigeria. *IOSR Journal of Research & Method in Education (IOS-JRME)*, *7*, 83.

Gimba, R. W. (2006). Effects of 3-dimensional instructional materials on the teaching and learning of mathematics among senior secondary schools in Minna metropolis. In *2nd SSSE Annual National Conference, Federal University of Technology, Minna. Held between 19th–2nd November*.

Hou, H. T., & Li, M. C. (2014). Evaluating multiple aspects of a digital educational problem-solving-based adventure game. *Computers in Human Behavior*, *30*, 29-38.

Hsu, T. C. (2017). Learning English with augmented reality: Do learning styles matter? *Computers & Education*, *106*, 137-149.

Ibili, E., & Sahin, S. (2015). The effect of augmented reality assisted geometry instruction on students' achievement and attitudes. *Teaching Mathematics and Computer Science*, *13*(2), 177-193.

Ibili, E., Çat, M., Resnyansky, D., Şahin, S., & Billinghurst, M. (2019). An assessment of geometry teaching supported with augmented reality teaching materials to enhance students' 3D geometry thinking skills. *International Journal of Mathematical Education in Science and Technology*, (In Press).

Ibili, E., Resnyansky, D., & Billinghurst, M. (2019). Applying the technology acceptance model to understand maths teachers' perceptions towards an augmented reality tutoring system. *Education and Information Technologies*, (In Press).

Jamali, S. S., Shiratuddin, M. F., Wong, K. W., & Oskam, C. L. (2015). Utilising mobile-augmented reality for learning human anatomy. *Procedia-Social and Behavioral Sciences*, *197*, 659-668.

Karadeniz, Ş. (2006). Design cues for instructional hypertext, hypermedia and multimedia, Yüzüncü Yıl Univesity Journal of Education, 3(1).

Kato, H., & Billinghurst, M. (1999). Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *Augmented Reality, 1999. (IWAR'99) Proceedings. 2nd IEEE and ACM International Workshop on* (pp. 85-94). IEEE.

Kaushik, D., & Jain, R. (2014). Natural user interfaces: Trend in virtual interaction. *arXiv preprint arXiv:1405.0101*.

Kılıç, E. (2007). The Bottle Neck in Multimedia: Cognitive Overload. *Gazi University Journal of Gazi Educational Faculty, 27(2),* 1-24.

Kılıç, E., & Karadeniz, Ş. (2014). Cinsiyet ve öğrenme stilinin gezinme stratejisi ve başarıya etkisi. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 24(3)*, 129-146.

Kimbrough, A. M., Guadagno, R. E., Muscanell, N. L., & Dill, J. (2013). Gender differences in mediated communication: Women connect more than do men. *Computers in Human Behavior*, *29*(3), 896-900.

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in psychology, 8*, 1997.

Lai, A. F., Chen, C. H., & Lee, G. Y. (2019). An augmented reality-based learning approach to enhancing students' science reading performances from the perspective of the cognitive load theory. *British Journal of Educational Technology*, *50*(1), 232-247.

Lawton, C. A., & Morrin, K. A. (1999). Gender differences in pointing accuracy in computer-simulated 3D mazes. *Sex roles*, *40*(1-2), 73-92.

Leahy, W., & Sweller, J. (2005). Interactions among the imagination, expertise reversal, and element interactivity effects. *Journal of Experimental Psychology: Applied*, *11*(4), 266.

Lee, E. A. L., & Wong, K. W. (2014). Learning with desktop virtual reality: Low spatial ability learners are more positively affected. *Computers & Education*, *79*, 49-58.

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., & Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior research methods*, *45*(4), 1058-1072.

Leue, M. C., Jung, T., & tom Dieck, D. (2015). Google glass augmented reality: Generic learning outcomes for art galleries. In *Information and Communication Technologies in Tourism 2015* (pp. 463-476). Springer, Cham.

Liou, Yang, Chen, & Tarng (2017). The influences of the 2d image-based augmented reality and virtual reality on student learning. *Journal of Educational Technology & Society*, *20*(3), 110-121.

Moro, C., Štromberga, Z., Raikos, A., & Stirling, A. (2017). The effectiveness of virtual and augmented reality in health sciences and medical anatomy. *Anatomical sciences education*, *10*(6), 549-559.

Nedim, S. (2013). The effect of augmented reality treatment on learning, cognitive load, and spatial visualization abilities. *Unpublished doctoral dissertation, University of Kentucky, Lexington, USA.*

Ong, C. S., & Lai, J. Y. (2006). Gender differences in perceptions and relationships among dominants of e-learning acceptance. *Computers in human behavior*, *22*(5), 816-829.

Paas, F., & Van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load, Learning and Instruction, 16(2), 87-91.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational psychologist, 38(1)*, 1-4.

Pantano, E., Rese, A., & Baier, D. (2017). Enhancing the online decision-making process by using augmented reality: A two country comparison of youth markets. *Journal of Retailing and Consumer Services*, *38*, 81-95.

Poushneh, A., & Vasquez-Parraga, A. Z. (2017). Discernible impact of augmented reality on retail customer's experience, satisfaction and willingness to buy. *Journal of Retailing and Consumer Services*, *34*, 229-234.

Robertson, J. (2012). Making games in the classroom: Benefits and gender concerns. *Computers & Education*, *59*(2), 385-398.

Sadi, O., & Lee, M. H. (2015). The conceptions of learning science for science-mathematics groups and literature-mathematics groups in Turkey. *Research in Science & Technological Education*, *33*(2), 182-196.

Safadel, P. (2016). Examining the Effects of Augmented Reality in Teaching and Learning Environments that Have Spatial Frameworks, (Unpublished Doctoral dissertation), In Department of Educational and Instructional Technology, Texas Tech University.

Shelton, B. E. (2003). How augmented reality helps students learn dynamic spatial relationships (Unpublished Doctoral dissertation, University of Washington), University of Washington.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*, *22*(2), 123-138.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and instruction*, *12*(3), 185-233.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. Educational psychology review, 10(3), 251-296.

Tabachnick, B. G., & Fidell, L. S. (2001). *Computer-assisted research design and analysis* (Vol. 748). Boston: Allyn and Bacon.

Venkadesh, V., & Morris, M. G. (2000). Why dont men ever stop to ask for directions. *Gender, social influence, and their role in technology acceptance and usage behaviour*.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, *46*(2), 186-204.

Wang, X., Truijens, M., Hou, L., Wang, Y., & Zhou, Y. (2014). Integrating Augmented Reality with Building Information Modeling: Onsite construction process controlling for liquefied natural gas industry. *Automation in Construction*, *40*, 96-105.

Wei, X., Weng, D., Liu, Y., & Wang, Y. (2015). Teaching based on augmented reality for a technical creative design course. *Computers & Education*, *81*, 221-234.

Weiser, E. B. (2001). The functions of Internet use and their social and psychological consequences. *CyberPsychology & behavior*, *4*(6), 723-743.

Wickens, C. D., & Hollands, J. G. (2000). Signal Detection, Information Theory, and Absolute Judgment. *Engineering psychology and human performance*, *2*, 24-73.

Wu, P. H., Hwang, G. J., Yang, M. L., & Chen, C. H. (2018). Impacts of integrating the repertory grid into an augmented reality-based learning design on students' learning achievements, cognitive load and degree of satisfaction. *Interactive Learning Environments*, *26*(2), 221-234.

Xie, H. (2003). Supporting ease-of-use and user control: desired features and structure of Web-based online IR systems. *Information processing and management*, *39*(6), 899-922.

Xue, H., Sharma, P., & Wild, F. (2018). User Satisfaction in Augmented Reality-Based Training Using Microsoft HoloLens, Computers, 8(1), 1-23.

Young, J. Q., Van Merrienboer, J., Durning, S., & Ten Cate, O. (2014). Cognitive load theory: Implications for medical education: AMEE guide no. 86. *Medical teacher*, *36*(5), 371-384.

7. APPENDIX

Table 5. Cognitive Load and Usability Scale item descriptions

Items		Items Descriptions	Mean	SD
Intrinsic Load	IL1	The topics covered during the lesson were very complex.	3.02	1.81
	IL2	The lesson covered formulas that I perceived as very complex.	3.29	1.91
	IL3	The lesson covered concepts and definitions that I perceived as very complex.	3.89	1.95
Extraneous Load	EL1	The instructions and explanations during the lesson were very unclear.	3.77	1.85
	EL2	The instructions and explanations during the lesson were full of unclear language.	4.29	1.86
	EL3	The instructions and explanations during the lesson were, in terms of learning, very ineffective.	4.26	1.91
Germane Load	GL1	The lesson really enhanced my understanding of the topics covered.	7.29	1.74
	GL2	The lesson really enhanced my understanding of the geometry.	7.37	1.75
	GL3	The lesson really enhanced my knowledge of concepts and definitions.	7.23	1.72
	GL4	The lesson really enhanced my knowledge and understanding of the subject.	6.82	1.57
Ease of Use	EU1	Using ARGTS is easy for me.	6.56	1.85
	EU2	My Interaction with the ARGTS is clear and understandable.	6.42	2.01
	EU3	I find it easy to get the ARGTS to do what I want it to do.	6.42	2.09
Usefulness	PU1	I find ARGTS to be useful to me	6.92	1.78
	PU2	Using ARGTS can improve my teaching performance	7.43	1.84
	PU3	Using ARGTS enables me to accomplish tasks more quickly.	6.90	1.80
Natural Interaction	NI1	The interaction interfaces in ARGTS have created the feeling of touching a real object.	5.76	1.00
	NI2	The interaction with user interfaces in ARGTS was similar to the users’ interaction with real-world objects.	6.17	2.11
	NI3	I felt a natural interaction with the virtual content in ARGTS.	5.98	1.99

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 396–414
https://dx.doi.org/10.21449/ijate.505863

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

Social Network Addiction Scale: The Validity and Reliability Study of Adolescent and Adult Form

Ibrahim Gokdas [1,*], **Yasar Kuzucu** [2]

[1] Department of Computer Education and Instructional Technology. Adnan Menderes University, Faculty of Education, 09010 Aydin / Turkey
[2] Psychological Counseling and Guidance in Education, Adnan Menderes University, Faculty of Education, 09010 Aydin / Turkey

ARTICLE HISTORY

Received: 31 December 2018
Revised: 19 July 2019
Accepted: 29 July 2019

KEYWORDS

Social network addiction scale,
Social media addiction,
Social media usage

Abstract: In this study, it was aimed to develop a valid and reliable social network addiction scale for adolescents and young adults. In the Exploratory Factor Analysis of the scale, the application was conducted to 425 high school students between 14-17 years of age and 310 young adults between 18-43 years of age. Confirmatory Factor Analysis was performed on a different group and for this purpose, 322 high school students and 197 young adults were included in the analysis. As a result of the analyses performed, the scale exhibited a-10-item and three-factor structure in both groups. The total variance explained was 71.51% for adolescents and 70.96% for young adults. The total Cronbach Alpha reliability coefficient of the scale was .87 for adolescents and .84 for young adults. With the 1st and 2nd level Confirmatory Factor Analysis performed on a similar study group, a good model was revealed for both adolescents and young adults. The Social Network Addiction Scale developed within the scope of this study is thought to have the adequate validity and reliability structure that can be used to measure social network addiction levels of adolescents and young adults.

1. INTRODUCTION

Depending on the widespread use of the internet and the developments in information technologies, social networks are getting into our lives increasingly day by day. Social networks where texting and sharing (photos, documents, videos, etc.) are performed intensively affect the lives of many people from different age groups with the opportunities they offer. Depending on this development process, the use of social network has become an increasingly popular free time activity in many countries (Kuss & Griffiths, 2011). Today, individuals tend towards social networks to participate in many different entertainment and social activities, including playing games, socializing, spending time, communicating and sending pictures (Allen, Ryan, Gray, Mclnerney, & Waters, 2014; Ryan, Chester, Reece, & Xenos, 2014). Such attractive

CONTACT: Ibrahim Gokdas ✉ igokdas@gmail.com Department of Computer Education and Instructional Technology. Adnan Menderes University, Faculty of Education, 09010 Aydin / Turkey

ISSN-e: 2148-7456 /© IJATE 2019

opportunities offered by social networks have an important role in the lives of many people from different age groups and affect their lives.

Social Network Websites that are defined as virtual communities where users can create individual and general profiles, interact with their friends and meet other people in line with common purposes (Kuss & Griffths, 2011) have made significant changes in the way people communicate with others (Vilca & Vallejos, 2015). Social networks, a new communication technology paradigm (Kang, Shin, & Park, 2013; LaRose, Connolly, Lee, Li, & Hales, 2014), have taken an important place in our lives with their popularity (LaRose, et al., 2014). According to 2017 data, the population of the world is 7,476 billion people and 3,773 of it are internet users. 2,789 billion people are actively using social networks. According to the usage ratio of 2016, the number of internet users has increased by 354 million with a 10% rise and the number of social media users has increased by 482 million with a 21% rise. Following the first Global Digital Report for internet usage in January 2012, the number of global users has increased more than 80% in five years (We are Social, 2017). When the annual change ratios are taken into account, the increase in the number of social network users is particularly noteworthy. A meta-analysis conducted found out that about 6% of the world's population has internet addiction (Cecilia & Yee-lam, 2014). This ratio corresponds to about 226 thousand of people when considered for 2017 data.

The increase in the ratio of users caused social networks to be considered as a normal modern phenomenon within the society (Boyd & Ellison, 2007). However, the increase in the time that people spend online on social networks (Kuss & Griffiths, 2011) has brought together the concerns about addiction and social network use (Andreassen, 2015; Griffiths, Kuss & Demetrovics, 2014). There is also increasing evidence that social network addiction is a mental problem that occurs in adolescents (Pantic, 2014; Ryan et al., 2014).

It is believed that the developments in the features of information technologies (laptop computers, tablet PCs, smartphones, etc.) have a significant role in the widespread use of social networking and the increase of addiction because new technologies support easy and fast access to social networking sites and it is known that excessive use of such new technologies can be addictive especially for adolescents (Echeburúa & de Corral, 2010). When the user profile of social networks is investigated, it is seen that especially adolescents are the most intensive user group (Van den Eijnden, Lemmens, & Valkenburg, 2016; Vilca & Vallejos, 2015). For example, when the data of 2017 is analyzed, it is noted that about 73% of Facebook users are the individuals between the age of 18-34, 9% of the users are between the age of 13-17and 10% of the users are between the age of 35-44 (We are Social, 2017). This intensity is particularly worrying in that the risk of social network addiction in especially adolescents and young adults has increased and it has caused adolescents to move away from the necessary activities for their improvement. (Park, Kim, & Cho, 2008).

The egocentric nature of social networks push people towards problematic use by contributing to the development of addiction behaviors. Similarly, social networks lead individuals to exhibit themselves different from what they really are and live delightful experiences (Kuss & Griffths, 2011). Furthermore, the opportunities offered by social networks make users happy (Choi & Lim, 2016; Yang, Liu, & Wei, 2016) and create excitement by filling a psychological gap in the lives of individuals (Echeburúa & De, 2010; Yang et al., 2016). Together with the popularity achieved by social networks and many benefits they provide the users (Kuss & Griffiths, 2012), spending too much time in a social network (Can & Kaya, 2016) is considered a sign of social network addiction (Gao, Liu, & Li, 2017; Turel & Serenko, 2012) and may cause psychological disorders (Salehan & Negahban, 2013).

Despite the fact that the latest edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) recognize Internet addiction as a temporary disorder in the appendix of this guide (APA, 2013), social network addiction does not still hold a status in DSM-5. The fact that social network addiction is not included in DSM-5 creates the impression that social network addiction is not a psychological problem (Van den Eijnden et al., 2016). However, there are also studies that do not support this situation (Pantic, 2014; Ryan, Chester, Reece, & Xenos, 2014). The fact that there are no explicit definitions and precautions for social network addiction affects doing researches about these widespread behaviors negatively (Van den Eijnden et al., 2016).

Different research suggests that excessive use of social networks is associated with anxiety, frustration, intolerance, anger, low self-esteem, impoverishment of social relationships, decrease in academic performance, verbal or physical aggression, and depression tendency (Cheung & Wong, 2011; Huang & Liang, 2009; Satici & Uysal, 2015). Besides, it was found that excessive use of social networks might lead to such negative conclusions as sleeping disorders (Dewald, Meijer, Oort, Kerkhof, & Bögel, 2010) and procrastination of sleeping time (Brunborg, Mentzoni, Molde, Myrseth, Skouverøe, Bjorvatn et al., 2011; Suganuma, Kikuchi, Yanagi, Yamamura, Morishima, Adachi et al., 2007).

Considering the psychological, social, economic, cultural and educational losses caused by social network addiction, it is significant to determine the level of social network addiction. However, the number of studies regarding social network addiction are insufficient (Kuss & Griffths, 2011; Andreassen, 2015). When the researches conducted are analyzed, it is possible to get the impression that Facebook addiction has the same meaning as social network addiction (Ryan et al., 2014; Griffiths, Kuss, & Demetrovics, 2014; Van den Eijnden, et al., 2016). Addiction scales developed in this respect are focused on Facebook addiction or problematic Facebook usage (Kuşay, 2013; Andreassen, 2015) and have become intense after 2011 (Ryan et al., 2014). Social networks exhibit different characteristics in terms of functionality and expediency. For example, social networks such as Blogger etc. for publishing content, YouTube, Slideshare etc. for sharing, Messenger, Skype etc. for chatting, Facebook, LinkedIn etc. for getting to know other people, Twitter, Twitpic etc. for expressing short ideas, Friendfeed, Foursquare etc. for sharing life are widely used (Kuşay, 2013). On the other hand, the ratio of people sharing on Instagram, Pinterest or Twitter instead of Facebook is increasing rapidly. Besides, the ratio of WhatsApp users from different age groups is also increasing rapidly. YouTube is a tool where especially adolescents watch and share videos. Today, the number of present social network websites is over 100 (Pantic, 2014). Lots of people from different age groups actively and intensely use more than one of these, not just one. Therefore, social networks have a strong social impact on the lives of their users.

In spite of the popularity of social network use among people, empirical studies analyzing the addiction to these networks are insufficient (Ryan et al., 2014). Considering this fact, it is important to have psychological tools to be able to identify early possible social network addiction (Vilca & Vallejos, 2015). However, the diversity in social networks makes the studies regarding social network addiction problematic. The first reason for this is the rapid change in the social network environment and the expansion of its interactive functions. This will cause the measurement tools targeting specific social networks to lose their up-to-datedness easily. The second reason is that the criterions that may cause social network addiction vary. These reasons will cause problems in the process in comparing the related researches carried out (Van den Eijnden et al., 2016). The distinctive nature of each social network environment and the differences in the opportunities it offers reveals that social network addiction should be considered as different from internet or Facebook addiction alone. Therefore, the development of studies on social network addiction requires the development and validation of a general social network addiction scale (Van den Eijnden et al., 2016).

When the literature was examined examined (Esgi, 2016; Fırat & Barut, 2018; Şahin, 2018; Tutgun-Ünal & Deniz 2015; Ülke, Noyan, & Dilbaz, 2017; Van den Eijnden et al., 2016), it was found that various scales had been developed regarding social network addiction in recent years. Insufficient number of researches in the field and the need for scale development in this regard have been the main problem. However, it can be seen that some of the measurement tools developed in Turkish language are directed to young adults and adults (Esgi, 2016; Tutgun-Ünal, & Deniz, 2015; Ülke, Noyan, & Dilbaz, 2017) while some others are directed to adolescents and young adults (Fırat & Barut, 2018; Şahin, 2018). Besides, the study of Taş (2017), who conducted the Turkish adaptation study of the short form of social media addiction scale developed by Van den Eijnden et al. (2016) for adolescents and young adults, also involves adolescents and young group. In the international field literature, no social network addiction scale involving a large target group was found. In this regard, Social Media Disorder Scale developed by Van den Eijnden et al. (2016) involves the group of 10-17 years of age. Bergen Social Media Addiction Scale (BSMAS) is a modified version of the previously approved Bergen Facebook Addiction Scale (BFAS) (Andreassen et al., 2012). In the scales, the word "Facebook" was replaced by the word "social media" and social media was defined as "Twitter, Instagram and etc.". Bergen Facebook Addiction Scale (Andreassen, Torsheim, Brunborg, & Pallesen, 2012) was later adapted as Bergen Social Media Scale (Andreassen, Pallesen, & Griffiths, 2017). The original scale involved university students.

When evaluated in general, it can be revealed that the scales in the literature for social network addiction differ from each other in terms of the target group. Furthermore, as the factor structures of the scales also differ from each (Esgi, 2016; Fırat & Barut, 2018; Şahin, 2018; Taş, 2017; Turgut-Ünal & Deniz 2015; Ülke, Noyan, & Dilbaz, 2017; Van den Eijnden et al., 2016), it is considered that they will be insufficient for the studies to be conducted and in providing comparability among different age groups. Therefore, there is a need for an easily applicable scale that involves a target population of wider age range.

Based on these basic justifications, the main purpose of the study was to develop a highly valid and reliable social network addiction measurement tool for adolescents, young adults and adults. The research is deemed important in terms of involving different target groups as adolescents, young adults and adults with regard to their social network addiction levels.

2. METHOD

2.1. Study Group and Process

Social Network Addiction Scale (SNAS) was implemented to 461 students between the ages of 15 and 18 from five different high schools studying in Efeler, the central district of Aydın province in the spring semester of 2016-2017 academic year. The participants were given the information that the data would not be considered personally and that there would not be only one correct answer for everyone. The application lasted about 15-20 minutes. However, because of faulty and missing information, and after excluding extreme values, 425 forms were included in the evaluation. In the first stage, Exploratory Factor Analysis (EFA) was performed on the data collected from 425 students. In order to test the results of EFA with Confirmatory Factor analysis (CFA), additional data was collected from 371 high school students of the same age group. Participant students were determined by convenience sampling and it was noted from each level of education that they were using at least one social network and were voluntarily participating. SNAS was also applied to the adult group. For the adult group, 750 people who had undergraduate education at Adnan Menderes University and who graduated from a higher education institution and participated in pedagogical formation training were asked to fill in the form via e-mail. The form of the scale was organized online. A total of 367 participants completed the scale during the three-week period. However, a total of 310 forms were included in the evaluation after excluding extreme values and EFA was performed on this dataset. In order to test the results of the EFA with CFA, 430 people were e-mailed and 201 of them responded. Nonetheless, 197 data were included in the evaluation after the extreme values were excluded. The distribution of the study groups is given in Table 1and Table 2.

Table 1. The distribution of the study group for the Exploratory Factor Analysis

	Adolescent (14-17 years of age)					Adult (18-45 years of age)			
	Applied		Valid			Applied		Valid/Returned	
Age	f	%	f	%	Education	f	%	f	%
14	98	21,3	91	21,4	Undergraduate	188	51.2	158	51
15	108	23,4	98	23,1	Graduate	179	48.8	152	49
16	153	33,2	140	32,9	Total	367	100	310	100
17	102	22,1	96	22,6	Age				
Total	461	100,0	425	100,0	18-22	172	46,5	135	43.5
					23-27	130	35,1	110	35.4
Gender					28-32	32	8,6	30	9.7
Male	125	27,1	114	26,8	33-37	23	8,6	23	7.5
Female	336	72,9	311	73,2	38 and above	12	3,8	12	3.9
Total	461	100,0	425	100,0		367	100	310	100
					Gender				
					Female	226	61,6	193	62.2
					Male	141	38,4	117	37.8
					Total	367	100	310	100

Table 2. The distribution of the study group for the Confirmatory Factor Analysis

	Adolescent (14-17 years of age)					Adult (18-45 years of age)			
	Applied		Valid			Applied		Valid/Returned	
Age	f	%	f	%	Education	f	%	f	%
14	84	22.6	73	24,2	Undergraduate	100	30,3	78	39,6
15	96	25.9	90	30,2	Graduate	330	69,7	119	60,4
16	150	40.4	126	31	Total	330	100	197	100
17	41	11	33	14,6	Age				
Total	371	100	322	100	18-22	19	9.6	19	9,6
					23-27	77	38.4	79	40,1
Gender					28-32	45	22.4	46	23,4
Male	148	39,9	117	36,3	33-37	23	11.6	23	11,7
Female	223	60,1	205	63,7	38 and above	30	18	30	15,2
Total	371	100,0	322	100,0	Total	201	100	197	100
					Gender				
					Female	113	56.2	110	55.8
					Male	88	43.8	87	44.2
					Total	201	100	197	100

2.1.1. The Development of Social Network Addiction Scale

In order to develop a measurement tool for social network addiction, literature was analyzed and questions were prepared taking into account especially the studies of Young (1998), Griffiths (2005), Block (2008), Tao (2010), and Van den Eijnden et al., (2016). By obtaining the views of three experts who had PhD degrees in the field of psychology and who studied on internet addiction, five inappropriate items with similar meanings were removed from the draft form. The remaining 30 items were taken into the trial form of the scale. All of the items are positive statements and they are in the 5-point- Likert form ranked as 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5 = Very Often. For the items in the trial form, EFA, CFA and item analysis were performed.

2.2. Data Collection Tools

In order to determine the criterion validity of the SNAS Adolescent Application, the relationship of the Problematic Mobile Phone Use Scale with the SNAS was analyzed. For the criterion validity of SNAS Adult Application, the relationship of it with the Internet Addiction Scale was analyzed.

2.2.1. Problematic Mobile Phone Use Scale

In order to determine the criterion validity of the SNAS, the Problematic Mobile Phone Use Scale developed by Augner & Hacker (2012) and adapted by Tekin, Güleş, & Çolak (2014) was utilized. Problematic Mobile Phone Use Scale is composed of three sub-dimensions in total. The first sub-dimension is defined as “addiction” (9 questions), the second sub-dimension is defined as “social relations” (7 questions), and the third sub-dimension is defined as “results” (10 questions). In the adaptation study, the three-factor scale explains 45% of the total variance. The Cronbach Alpha value of the scale was found 0.85. Besides, the Cronbach Alpha value of the first sub-dimension (addiction) was found 0.73, that of the second sub-dimension (social relations) was found 0.60, and that of the third sub-dimension(results) was found 0.85 (Tekin et al., 2014).

2.2.2.Internet Addiction Scale

The Internet Addiction Scale developed by Young (1998) and adapted to Turkish by Cakir and Horzum (2008) was used to determine the criterion validity of the scale. The Turkish adaptation of the scale is composed of three sub-dimensions in total. The first sub-dimension is defined as “preferring being online to daily life” (8 items), the second sub-dimension is defined as “having desire to increase the duration of being online” (7 items), and the third sub-dimension is defined as “the problems arising from being online” (4 items)". The total variance explained was 52.83% and the total Cronbach Alpha internal consistency coefficient was found .90 (Çakır & Horzum, 2008).

2.3. Data Analysis

SPSS 22.0 (SPSS Inc.) and LISREL 8.80 (Joreskog & Sorbom, 1993) statistical package programs were used in the analysis process. The data of the scale applied separately to adolescents and adults were analyzed using EFA and CFA techniques for the construct validity. By examining the measurement invariance in adolescent and adult samples, it was tested whether the measurement tool was appropriate for the comparisons between groups. Furthermore, item-test score correlations, test-retest scores correlation, internal consistency McDonald Omega coefficient (McDonald, 1999) were calculated. T-test was performed to test whether the items of the scale distinguished between the lower and upper 27% groups. Item Response Theory (IRT) was used to check the reliability results obtained.

3. FINDINGS

3.1.Adolescent Application

3.1.1. Pre-analyses

In order to determine whether the data showed normal distribution or not, Skewness and kurtosis values were examined. Skewness was found .61 and Kurtosis was found -.03. The fact that both values are between the range of -1, +1 implies that they show normal distribution. In addition to Skewness and Kurtosis analyses, Kolmogorov-Smirnov test results (p> .05) support normal distribution.

Kaiser Meyer Olkin (KMO) coefficient was used to determine whether the data structure was appropriate for factor analysis in terms of the sample size of the SNAS adolescent application. As a result of the analysis, KMO value was determined as 0.87. The fact that KMO value is high means that each variable in the scale can be estimated well by the other variables (Çokluk, Şekercioğlu, & Büyüköztürk, 2012). Another indicator for the appropriateness of the data for factor analysis is the Anti-image Correlation Matrix. These values need to be above 0.5 and the values below this must be excluded from the analysis (Field, 2009). The diagonal values for each variable in the anti-image matrix vary between .80 and .91. The fact that all the values at the intersection point are above 0.5 indicates that it is accurate to include all the items in the scale.

3.1.2. The Validity of Social Network Addiction Scale Adolescent Application

After determining that the sample size is appropriate for factor analysis, the factor structure for the construct validity of the scale was determined by performing EFA. The purpose of performing EFA is to gather the variables that are related to each other and that measure the same quality together, and to reduce the number of items forming the scale (Aksu, Eser, & Güzeller, 2017). CFA was performed to test whether the restricted structure defined by EFA was verified as a model (Çokluk, Şekercioğlu, & Büyüköztürk, 2012).

After the first factor analysis with a total of 30 items, the items were collected in 5 sub-dimensions, with eigen values greater than 1. However, the items numbered 1, 3, 4, 7, 8, 10, 11, 12, 13, 14, 15, 18, 19, 20, 21, 22, 24, 25, 26, 27 with factor loadings below 0.30 and that were overlapping were gradually removed from the scale. Factor analysis was made again by removing one item at each step. As a result, 20 items were removed from the scale and the remaining 10 items were collected in 3 sub-dimensions. The items of each sub-dimension were examined and it was determined that they were grouped under the factor to which they were related. To clarify the relationship among factors, direct oblimin rotation (the oblique rotation technique of Principal Component Analysis) was used. As a result of the EFA performed by using the oblimin rotation method, it was found that the eigenvalue of the first factor was 4.6 and the variance it explained was 26.27, the eigenvalue of the second factor was 1.49 and the variance it explained was 25.61, the eigenvalue of the third factor was 1.05 and the variance it explained was 19.63. The total variance explained by the scale was found 71.51%. When the eigenvalues and cumulative variance percentages of the three factors were taken into consideration, it was determined that the scale had three factors. The findings obtained as a result of the EFA performed for SNAS Adolescent Application revealed that the construct validity of the scale was sufficient. The factors formed after EFA and the items collected under each factor are given in Table 3.

When Table 3 is examined, it can be seen that the first factor is composed of 4 items (2, 5, 6, 9), the second factor is composed of 3 items (28, 29, 30) and the third factor is composed of 3 items (16, 17, 23). The results reveal that each item is clustered under a factor that is related with a value that is more than twice as much as the factor loading value that they have in other factors. This finding, which shows that the items differentiate in terms of factors, support the construct validity of the scale.

When the scree plot conducted to reveal the factor structure of the scale is analyzed, it can be seen that the graph curve shows a sharp decrease till the third factor and that the curve proceeds horizontally after the third factor (Figure 1). This finding supports the three-factor structure of the scale.

Table 3. Factor Loadings of SNAS Adolescent Application

Factors	Items		Factor Loadings		
			1	2	3
Control Difficulty	Q1	I find myself surfing the social networks in most of my daily life. (5)	.890	.030	.115
	Q2	I spend most of my time in social networks. (9)	.807	.034	.009
	Q3	I do not give up using social networks even if they affect my daily life.(6)	.765	.036	.148
	Q4	I make an effort to use social networks every day. (2)	.716	.128	.126
Negativeness in Social Relations	Q8	I feel happy to share my ideas on social networks. (23)	.111	.887	.017
	Q9	I prefer to share my daily activities on social networks. (16)	.086	.768	.082
	Q10	I express myself better on social networks. (17)	.082	.685	.097
Decrease in Functions	Q5	The time I allocate for my work/lessons have decreased since I began to use social networks. (29)	.013	.009	.911
	Q6	My performance at work/school have decreased since I began to use social networks. (28)	.013	.026	.909
	Q7	I have begun to have problems focusing on my work/school since I began to use social networks. (30)	.037	.016	.875

Notes: N =425 * p<.05 **p<.01

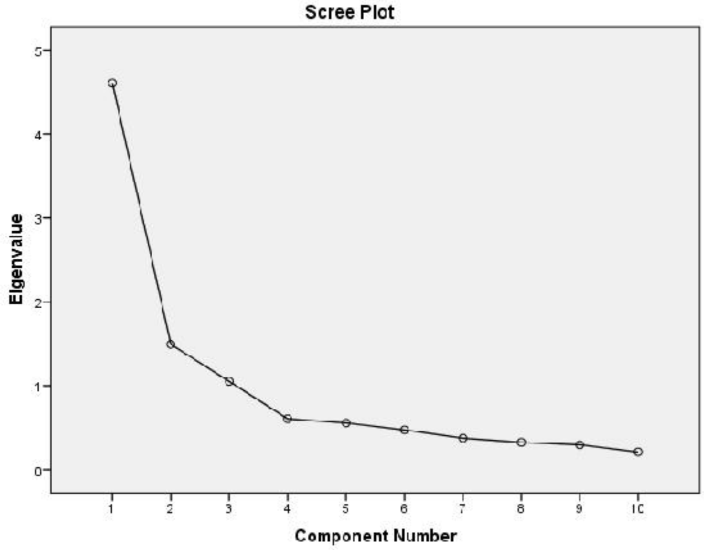


Figure 1. SNAS Adolescent Application Scree Plot Graph

Following this phase, the items in each sub-dimension were examined as a whole and a factor structure consistent with the theoretical framework was observed. Within this context, in relation to the literature on addiction, the first sub-dimension of the scale was named as “Decrease in Functions”, the second sub-dimension was named as “Control Difficulty” and the third sub-dimension was named as “Negativeness in Social Relations”. In order to determine whether there were significant correlations among the factors forming SNAS Adolescent Application, Pearson Product-Moment Correlation Analysis was performed. It was revealed that the relationship of “Control Difficulty” factor with “Decrease in Functions” and “Negativeness in Social Relations” factors was found as .52 and .41, respectively; and the relationship between “Decrease in Functions” and “Negativeness in Social Relations” was determined as .30. The results obtained, consistent with the literature (Şahin, 2018), show that there was a positive significant relationship among all the sub-dimensions of the scale *p*≤.001.

First level and second level Confirmatory Factor Analysis (CFA) was performed to evaluate the applicability of the three sub-dimensions of SNAS Adolescent Application to the data obtained from the study group. The models obtained from these analyses are given in Figure 2 and Figure 3.

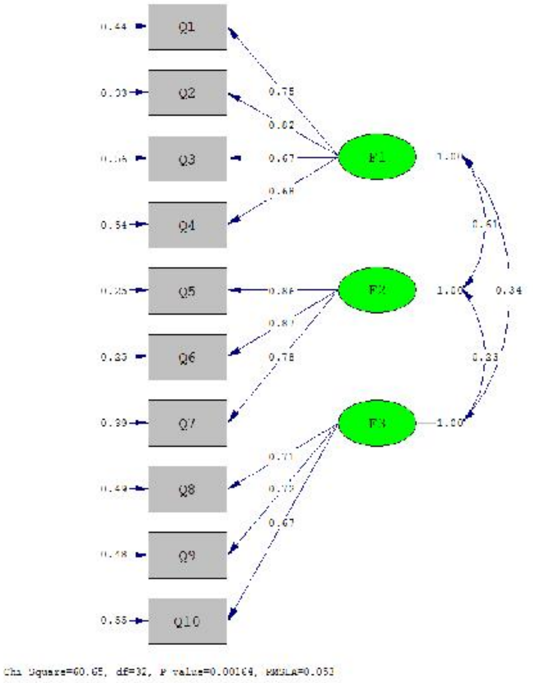


Figure 2. SNAS Adolescent Application 1st Level CFA

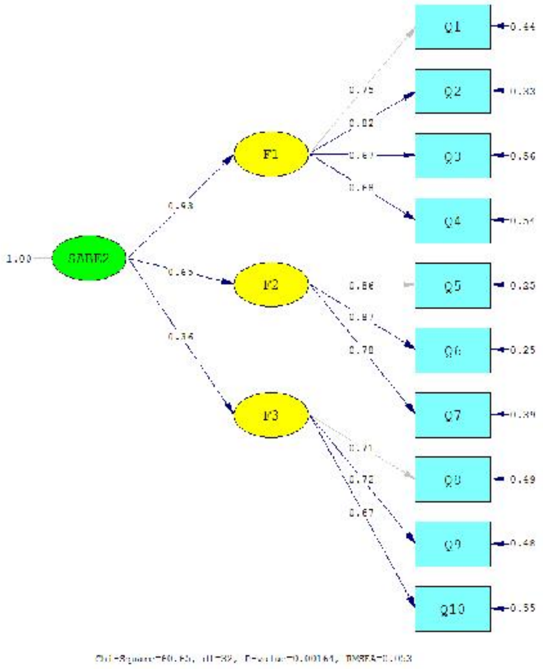


Figure 3. SNAS Adolescent Application 2nd Level CFA

First and second level CFA was performed for the 10-item structure that was collected under three factors as a result of EFA performed for SNAS adolescent application. When the findings revealed as a result of CFA were evaluated, χ2/sd ratio for the first and second level was determined as 2.08 (χ2/sd=66.65/32). The fact that χ2/sd ratio obtained as a result of first and second level CFA is between 2.0≤3.0 correspond to an acceptable fit. RMSEA fit index value was determined as 0.053 as a result of first and second level CFA. The fact that RMSEA fit index value is below 0.08 can be interpreted as acceptable fit (Kline, 2015). It was determined that, among the fit index values related to the model as a result of the first and second level CFA, AGFI was 0.94, GFI was 0.96, standardized RMR fit index value was 0.041, NFI fit index value was 0.97, and CFI fit index value was 0.99. When all the values related to data fit of the model are taken into consideration, it can be seen that the model formed shows adequate level of fit with the data.

An additional CFA was performed to support the multifactorial structure of SNAS Adolescent Application; the results of first and second level factor analysis were compared with the 1-factor analysis of the scale. Scale was assumed one dimensional and it produced following statistics: χ2/sd ratio of the fit values used in the model comparisons was calculated as 16.5 (χ2/sd=580.1/35, NFI=0.77, GFI=0.73, CFI=0.78 RMSEA=0.22). The results obtained showed that the 1-factor structure had poorer fit values than the 3-factor structure. In order to determine the criterion validity of SNAS Adolescent Application, the relationship between Problematic Mobile Phone Use Scale (PMPUS) and SNAS Adolescent Application was examined with Pearson Product-Moment Correlation Analysis and it was found that there was a positive (r=.55) and statistically significant (p≤.001) relationship between the two variables.

3.1.3. The Reliability of SNAS Adolescent Application

Item analysis was conducted to determine the contribution of the items in the scale to the implicit structure they belong to, and to measure the level of discrimination between the items with and without relevant characteristics in the structure they belong to (Erkuş, 2012). It was revealed that item total correlation coefficients varied between .41 and .73. On the condition that the items are congeneric measurements, McDonald Omega coefficient is used (McDonald, 1999). McDonald Omega coefficient of the overall scale was calculated as .87. The reliability analysis for each factor of the scale was also conducted. As a result of the analysis performed for this purpose, it was found for the first factor that McDonald Omega coefficient was .76 and item total correlation coefficients varied between .61 and .72. For the second factor, McDonald Omega coefficient was found .81 and item total correlation coefficients varied between.51 and .56. For the third factor, McDonald Omega coefficient was found .72 and item total correlation coefficients varied between .77 and .83. It can be seen that the reliability values of the overall and sub-dimensions of the SNAS Adolescent Application are generally acceptable values for social sciences.

It was also analyzed whether there was a significant difference between the individuals with low scores and high scores. As a result of the t test conducted to determine the difference between the responses of the individuals in the lower 27% group and the responses of the individuals in the upper 27% group to all the items in the scale, the items' *t* values varied between 4.14 (p<.001) and 10.67 (p<.001) and a significant difference was found. In the analysis performed, it was found that the variances were heterogeneous.

Item Response Theory (IRT) was used to confirm the reliability results obtained. IRTPRO 4.2 software was used to analyze with IRT. For this purpose, the three-factor structure fit of the scale was analyzed by using the two-parameter logistic model (2PL) to examine the items. The item difficulty (a) and item discrimination power (b) that were considered important were analyzed according to this formulation (Hambleton, Swaminathan & Rogers, 1991). Besides, X^2value, which is the measure of item-model fit, and the items that were insignificant (p<= 0,01) were also examined. The calculated "a" and "b" item parameter values and X^2 values are given in Table 4.

Table 4. Parameter values in terms of SNAS Adolescent Application according to IRT

Item	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*	b_4	*s.e.*	X^2	*df*	*p*
1	1.39	0.17	-2.69	0.33	-0.90	0.15	0.79	0.13	2.45	0.27	57.51	57	0.457
2	1.62	0.20	-2.33	0.26	-0.37	0.11	1.25	0.14	3.00	0.33	33.92	49	0.950
3	1.41	0.17	-1.58	0.20	0.06	0.11	1.31	0.16	3.59	0.44	50.42	58	0.750
4	1.38	0.17	-1.41	0.19	0.04	0.11	0.90	0.13	1.97	0.22	60.59	67	0.697
5	2.82	0.39	-0.66	0.09	0.11	0.09	0.92	0.12	1.67	0.17	67.38	51	0.061
6	2.79	0.42	-0.46	0.08	0.45	0.10	1.22	0.15	2.01	0.21	102.04	49	0.058
7	2.36	0.30	-0.55	0.09	0.46	0.10	1.38	0.16	2.24	0.24	70.02	49	0.045
8	0.65	0.12	-1.65	0.64	2.11	0.76	4.84	1.62	8.13	2.74	90.14	79	0.183
9	0.47	0.12	-0.45	0.28	2.16	0.58	4.87	1.26	7.37	1.95	82.48	70	0.145
10	0.64	0.13	-0.52	0.22	1.22	0.28	3.39	0.65	5.51	1.10	76.31	64	0.139

The item discrimination parameter provides information about the quality of the item. While items with A parameter value below 0.5 are regarded as weak in terms of discrimination (De Beer, 2004), those above 1 are not deemed adequate (Gültaş, 2014). From Table 4, it can be seen that all the values except for the value of item 9 are sufficient and the values of item 8 and 10 are at the borderline. Item discrimination serves to differentiate between the individuals with low and high social network addiction. Item difficulty indicates where the item is functional on the social network addiction level of the item. High level of "b" value exhibits that the item is functional or it measures among the individuals with high addiction levels, whereas low level of "b" value indicates that the item is functional or it measures among the individuals with low addiction levels. Item difficulty value varies between -2.69 and 8.14. It was noted that while the first threshold value of the scale (Likert 1 and 2 interval) was about -2, the second threshold value was 0, the third threshold value was 1, and the fourth threshold value varied between 2 and 8. This suggests that the scale is better discriminated in the individuals with high social network addiction. It was found that from the X^2values indicating item model fit, only item 7 was significant and did not meet the model fit. This item was not removed from the scale due to the fact that it had one of the highest factor loadings with a factor loading of .84, and that its item total correlation was high (.60) as a result of EFA.

3.2. SNAS Adult Application

3.2.1. Pre-analyses

KMO coefficient was used to determine whether the data structure was appropriate for factor analysis in terms of the sample size of the SNAS adult application. As a result of the analysis, KMO value was determined as 0.84. Besides, the Anti-Image Correlation Matrix intersection values were also analyzed and it was found that these values varied between .78 and .91. As the values at this intersection point were above 0.5, it was determined that it was accurate to include all the items in the scale.

In order to determine whether the data showed normal distribution or not, Skewness and kurtosis values were examined. Skewness was found .52and Kurtosis was found -06. The fact that both values are between the range of -1, +1 implies that they show normal distribution. Kolmogorov-Smirnov test results (p>.05) also support normal distribution.

3.2.2. The Validity of SNAS Adult Application

EFA and CFA were performed for the construct validity of the scale. After the first factor analysis with a total of 30 items, the items were collected in 5 sub-dimensions, with eigenvalues greater than 1. However, the items numbered 1, 3, 4, 7, 8, 10, 11, 12, 13, 14, 15, 18, 19, 20, 21, 22, 24, 25, 26, 27 with factor loadings below 0.30 and that were overlapping were gradually removed from the scale. Factor analysis was made again by removing one item at each step. As a result, 20 items were removed from the scale and the remaining 10 items were collected in 3 sub-dimensions. The items of each sub-dimension were examined and it was found that they were grouped under the factor to which they were related. To clarify the relationship among factors, direct oblimin rotation (the oblique rotation technique of Principal Component Analysis) was used. Within this context, it was determined that the first factor explained 26.2%, the second factor explained 25.26% and the third factor explained 19.5% of the total variance. The total variance explained by the scale was found 70.96%. The findings obtained as a result of the factor analysis performed for SNAS Adult Application reveal that the validity of the scale was sufficient. In addition, it was determined that it had the same factor structure with the adolescent application. The factors formed after EFA for SNAS Adult Application and the factor loadings are given in Table 5.

As can be seen in Table 5, the first factor is composed of three items (28, 29, 30) and the factor loadings vary between .88 and .90. The second factor is composed of four items (2, 5, 6, 9) and the factor loading values vary between .71 and .81. The third factor is composed of three items (16, 17, 23) and the factor loadings vary between .76 and .80.

Table 5. Factor Loadings of SNAS Adult Application

Factors	Items		Factor Loadings		
			1	2	3
Control Difficulty	Q5	I make an effort to use social networks every day.(2)	.842	.023	-.167
	Q4	I find myself surfing the social networks in most of my daily life.(5)	.817	.057	.050
	Q6	I spend most of my time in social networks. (9)	.737	.022	.150
	Q7	I do not give up using social networks even if they affect my daily life. (6)	.692	.020	.186
Negativeness in Social Relations	Q8	I feel happy to share my ideas on social networks.(23)	.075	.826	.008
	Q9	I express myself better on social networks.(17)	.001	.792	.080
	Q10	I prefer to share my daily activities on social networks.(16)	.107	.766	.095
Decrease in Functions	Q1	My performance at work/school have decreased since I began to use social networks. (28)	.049	.011	.925
	Q2	I have begun to have problems focusing on my work/school since I began to use social networks.(30)	.038	.043	.898
	Q3	The time I allocate for my work/lessons have decreased since I began to use social networks.(29)	.097	.012	.878

Notes: N=310 * p<.05 **p<.01

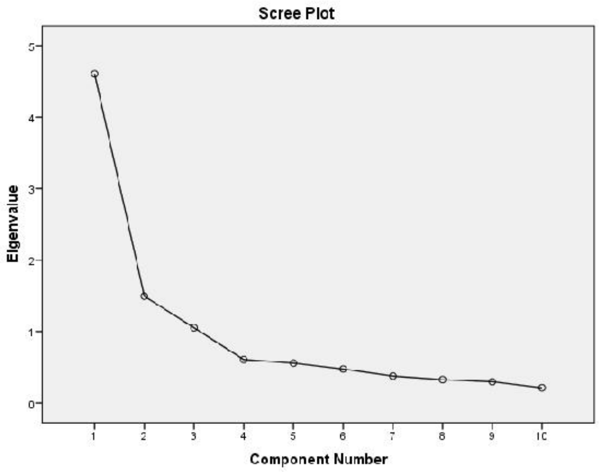


Figure 4. SNAS Adult Application Scree Plot Graph

When the "Scree Plot" graph is examined, it can be seen that the curve shows a sharp decrease till the third factor and that the curve proceeds horizontally after the third factor (Figure 4). The results are consistent with the previous results showing that the scale has a three-factor structure. After this process, it was analyzed whether there were any significant relationships between the factors forming the scale. As a result of Pearson Product-Moment Correlation Analysis conducted to test whether there was a significant relationship among the sub-dimensions of the scale, consistent with the literature (Andreassen, 2012; Esgi, 2017; Şahin, 2018; Şahin & Yağcı, 2017; Ülke et al., 2017), it was found that there were positive significant relationships among all the factors of the scale (p<.001). It was determined that the relationship of "Control Difficulty" factor with "Decrease in Functions" and "Negativeness in Social Relations" factors we as .40 and .47, respectively, and the relationship between "Decrease in Functions" and "Negativeness in Social Relations" was .23.

First level CFA was performed to determine whether the 10-item, 3-factor structure of the scale achieved after EFA performed for SNAS Adult Application would be verified.

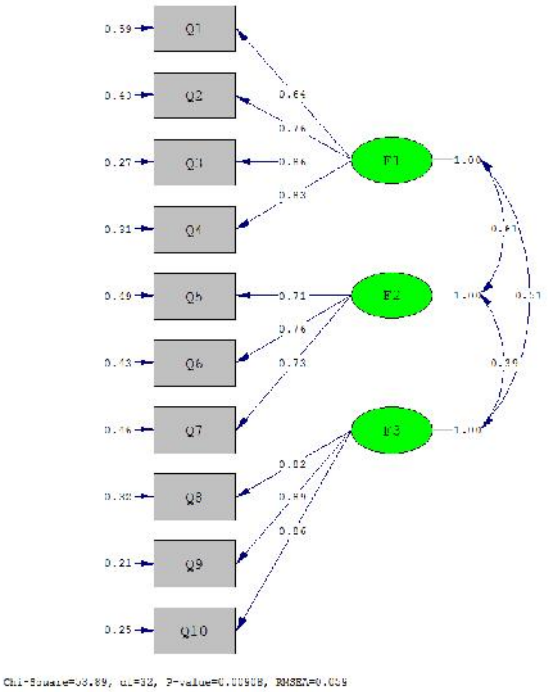


Figure 5. SNAS Adult 1stLevel CFA

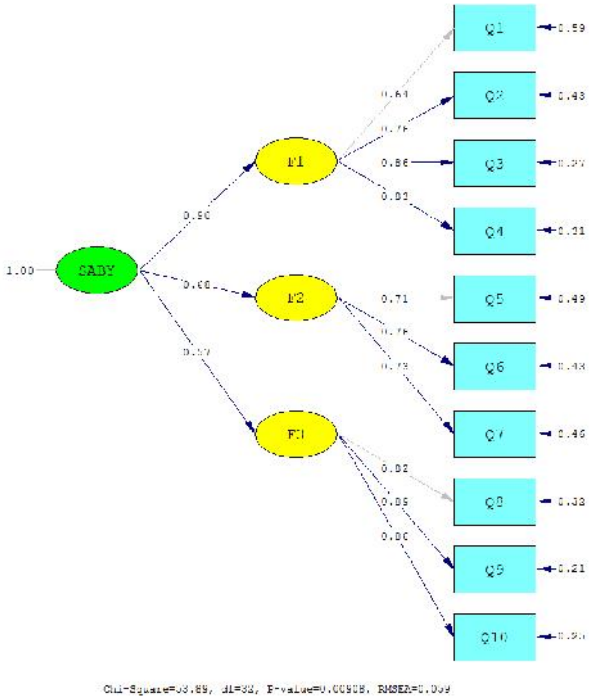


Figure 6. SNAS Adult 2ndLevel CFA

As a result of the first level (Figure 5) and second level (Figure 6) CFA performed for SNAS Adult Application, χ2/sd ratio was calculated as 1.68 (χ2/sd=53.89/32) and these values correspond to good fit. It was determined that, of the fit indexes, AGFI was .91, GFI was .95, standardized RMR fit index value was .056, NFI fit index value was .97, and CFI fit index value was .99. RMSEA fit index value for both levels was found as .059. When all the values related to data fit of the model are considered, it can be seen that the model formed shows adequate level of fit with the data.

To compare 1-factor and multifactorial structure, an additional CFA was performed Scale was assumed one dimensional and the obtained fit values ((χ2/sd=440.25/35=12.5, NFI=0.76, GFI=0.69, CFI=0.78, RMSEA=0.24) indicated that the 1-factor structure had poorer fit values than the 3-factor structure. This result supports to the multifactorial structure of SNAS Adult Application.

In order to determine whether the properties of the scale are invariant in different groups, measurement invariance was examined. While the measurement invariance of the factor structure of the scale was being measured for the adolescent and adult sample, multiple-group confirmatory factor analysis was used. For this purpose, 4 hierarchical models; structural invariance, metric invariance, strong invariance and strict invariance, which are commonly used in the literature were tested. In this study, it was examined whether the invariance conditions of ΔCFI ≤ -0.01 for multiple group confirmatory factor analysis study files which are compatible with the data of were obtained. The fact that ΔCFI value obtained as a result of the comparison of the two models is equal to -.01 or below can be used as the evidence that the measurement equivalence is achieved (Wu et al., 2007).

The findings regarding the invariance steps tested are present in Table 6. "The Structural Invariance Model" in the table represents the factor loads, regression constant and the error variances free model; "The Weak Invariance Model" in the table represents the factor loads constant, regression constants and error variances free model; "The Strong Invariance Model" in the table represents the factor loads, regression constants and error variance free model; and

“The Strict Invariance Model” in the table represents the factor loads, regression constants and error variances constant model.

Table 6. Fit statistics regarding measurement invariance

Steps	χ2	d	CFI	GFI	RMSEA	ΔCFI
Structural Invariance	87.84	67	.99	.97	.031	-
Weak (Metric) Invariance	44.76	32	1.00	1.00	.036	0.01
Strong (Scalar) Invariance	44.76	32	1.00	1.00	.036	0.01
Strict Invariance	87.84	87	1.00	.97	.006	0.01

As can be seen in Table 6, the fit indexes obtained as a result of multi-group CFI and ΔCFI values obtained as a result of CFI difference test can be interpreted for each step as follows. According to the results, it is seen that the structural invariance is provided and this finding shows that the measured structures use the same conceptual perspectives in responding to the scale items of the adolescents and adults. The finding regarding the metric invariance indicates that the factor structures of the variables taken in the model are the same in the adolescent and adult groups. It is confirmed that the strong invariance is provided and the constant number in the regression equations formed for the items is invariant between the groups. In the last stage, considering the ΔCFI value calculated with the fit indexes, it is accepted that the error terms regarding the items forming the measurement tool are invariant between the comparison groups. Hierarchical analysis results, factor structure and pattern of the scale, factor loads, regression constants, and error variances are seen to be invariant for the adolescent and adult groups.

For the criterion validity of SNAS Adult Application, the relationship with Internet Addiction Scale was examined. As a result of Pearson Product-Moment Correlation Analysis performed for the criterion validity, it was determined that there was a positive (r=.65) and statistically significant relationship (p≤.001) between the scales.

3.2.3. Reliability Studies

Reliability analyses were performed both for the overall and for the factors of SNAS Adult Application. On the condition that the items are congeneric measurements, McDonald Omega coefficient is used (McDonald, 1999). McDonald Omega coefficient of the overall scale was calculated as .91. McDonald Omega value for the first factor was .83; item total correlation coefficients varied between .78 and .84. For the second factor, McDonald Omega value was .76; item total correlation coefficients varied between .54 and .73. For the third factor, McDonald Omega value was .91; item total correlation coefficients varied between .52 and .55. As all the values in the reliability analysis both for the overall and for the factors of SNAS Adult Application are above 0.70, it can be said that the reliability of the scale is high. In the reliability analysis for the 10 items included in the scale, the item total correlation coefficients of the items varied between .37 and .66.

Item analysis was performed to determine whether there was a difference between the responses of the individuals with low scores (the lower 27% group) and high scores (the upper 27% group). As a result of the *t* test performed for this purpose, it was observed that *t* values of the items varied between 6.09 (p<.001) and21.03 (p<.001) and there was a significant difference.

Within the framework of Item Response Theory, item difficulty (a), item discrimination power (b) and item-model fit (X^2) was examined and the results are given in Table 7.

When Table 7 is analyzed, the fact that all the values are above 1 reveals that the items have good level of discrimination. High level of "b" value exhibits that the item is functional or it measures among the individuals with high addiction levels, whereas low level of "b" value indicates that the item is functional or it measures among the individuals with low addiction levels. Item difficulty value varies between -2.32 and 3.26. It was noted that while the first threshold value of the scale (Likert 1 and 2 interval) was between 0 and 1, the second threshold value varied between 0 and 1, the third threshold value varied between 1 and 2, and the fourth threshold value varied between 2 and 3. This suggests that the scale is better discriminated in the individuals with high social network addiction. The fact that all the X^2values indicating item model fit are insignificant shows that all the items meet the model fit.

Table 7. Parameter values in terms of SNAS Adult Application according to IRT

Item	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*	b_4	*s.e.*	X^2	*df*	*p*
1	1.53	0.22	-2.32	0.33	-1.35	0.22	-0.15	0.14	0.59	0.15	54.65	45	0.153
2	2.08	0.28	-0.77	0.16	0.09	0.12	1.11	0.14	1.72	0.19	53.93	44	0.144
3	2.43	0.34	-1.14	0.16	-0.38	0.12	0.40	0.11	1.45	0.16	60.30	44	0.051
4	3.00	0.43	-0.61	0.13	0.29	0.11	0.97	0.12	1.71	0.17	48.88	38	0.110
5	1.27	0.20	-1.01	0.22	0.78	0.16	1.96	0.28	3.26	0.51	49.03	45	0.314
6	1.35	0.22	-0.21	0.16	0.83	0.16	1.92	0.27	3.10	0.47	51.97	46	0.252
7	1.01	0.18	-0.82	0.23	0.65	0.19	1.91	0.33	3.20	0.56	70.70	52	0.053
8	1.73	0.35	0.66	0.13	1.76	0.23	2.38	0.34	3.16	0.54	33.63	30	0.295
9	1.63	0.31	0.21	0.13	1.34	0.18	2.35	0.33	2.91	0.45	52.26	37	0.049
10	1.70	0.33	0.47	0.12	1.35	0.18	2.07	0.28	2.83	0.43	44.54	39	0.249

4. DISCUSSION and CONCLUSION

The presence of social network addiction can be discussed depending on the definition of addiction used. However, there are evidences that some social network users are experiencing addiction-like symptoms due to excessive use. Besides, many studies have revealed that social networks are addictive (eg. Echeburúa & de Corral, 2010; Grffiths, Kuss, & Demetrovics, 2014; Pantic, 2014; Ryan et al., 2014). It can be seen in the literature that the researchers investigating social network addiction focus primarily on Facebook addiction (Andreassen, 2015). However, it has been discussed that Facebook is just a social network and therefore, there is a need for valid scales involving other social network sand measuring social network addiction (Griffiths et al., 2014). Although social networks, which can be considered as the sub-dimension of the internet, have some similar characteristics in terms of their intended use, they differ in the uses specific to individual and purpose (Kuşay, 2013; Van den Eijnden et al., 2016).

When the literature was examined, it could be seen that the scales developed differed in terms of factor structures and target groups and the total variance range explained varied between 35% and 59% (Esgi, 2016; Fırat, & Barut, 2018; Şahin, 2018; Şahin, & Yağcı, 2017; Taş, 2017; Tutgun-Ünal, 2015; Tutgun-Ünal, & Deniz, 2015; Ülke, Noyan, & Dilbaz, 2017; Van den Eijnden et al., 2016). The factor structures of these scales, which were developed for different age groups differed from each other. On the other hand, the factor structure of the social media addiction scale, which was developed by Bakır Ayğar & Uzun (2018) and whose target group was university students, was similar to the factor structure of SNAS. As could be seen, each scale was structured according to different age groups and their factor structures differed from each other. Besides, for the criterion validity in the measurement tools developed to measure social network addiction, Bakır Ayğar & Uzun (2018) used the problematic internet use scale; and Van den Eijnden et al. (2016) used compulsive internet use scale. In both studies, it was determined that the correlation with the scale used for the criterion validity was high. The criterion validity was not examined in the other scales developed for social network addiction (Eşgi, 2017; Şahin, 2017; Şahin & Yağcı, 2016).

The scales in the literature are generally dispersed in terms of target groups and factor structures. This situation is thought to have a negative effect on the comprehensive comparability between the developmental periods regarding social network addiction. SNAS developed in this study is significant in terms of sorting out this problem.

The factor structures of SNAS, which was composed of a total of 10 items, were named as "Control Difficulty", "Decrease in Functions" and "Negativeness in Social Relations".

As a result of the factor analysis performed, the total variance explained and factor load values were high in adolescent and adult form. The first level CFA and second level CFA results of the scale revealed that the model showed adequate fit with the data. McDodalds Omega reliability coefficient values for each sub-dimension, which was conducted to determine the reliability of the scale, also showed that the scale was reliable. Furthermore, IRT was used to confirm the reliability results obtained. As a result of this analysis, it was found that only the item number 9 in the adolescent form was weak in terms of discrimination. However, due to the fact that it was too close to the acceptable value and that both the factor load value and the total correlation coefficient were high, the item was not excluded from the scale. As a result of Pearson Product Moment Correlation Analysis performed for the criterion validity of adolescent and adult forms, a positive and statistically significant relationship was found between the scales. The highest score that can be obtained from the scale is 50 and the lowest score is 10. As the score obtained from the scale goes up to 50, addiction level increases, too. The scale developed involves adolescents, young adults and adults between 14-45 years of age.

The scale has great power of explaining the variable it intends to measure with a small number of well-working items. With this feature, the scale will provide researchers convenience and flexibility with the researches targeting different age groups and for possible comparisons.

Orcid

İbrahim GÖKDAŞ https://orcid.org/0000-0001-7019-8735
Yaşar KUZUCU https://orcid.org/0000-0002-8487-9993

5. REFERENCES

Aksu, G., Eser, M. T., & Güzeller, C. O. (2017). *Açımlayıcı ve Doğrulayıcı Faktör Analizi ile Yapısal Eşitlik Modeli Uygulamaları*. Ankara: Detay yayıncılık.

Allen, K. A., Ryan, T., Gray, D. L., Mclnerney, D. M., & Waters, L. (2014). Social media use and social connectedness in adolescents: The positives and the potential pitfalls. *The Australian Educational and Developmental Psychologist, 31*, 18 - 31. https://doi.org/10.1017/edp.2014.2

Andreassen, C. S. (2015). Online social network site addiction: A comprehensive review. Technology and Addiction (M Griffiths, Section Editor). *Current Addiction Reports. 2*, 175–184. DOI 10.1007/s40429-015-0056-9

Augner, C., & Hacker, G.W. (2012). Associations between problematic mobile phone use and psychological parameters in young adults. *International Journal of Public Health*. *57*(2), 437-41. DOI: 10.1007/s00038-011-0234-z

APA (American Psychiatric Association). (2013). Diagnostic and Statistical Manual of Mental Disorders (DSM-5). *American Psychiatric Association Publishing*.

Ayğar Bakır, B. & Uzun, B. (2018). Sosyal Medya Bağımlılığı Ölçeği'nin geliştirilmesi: Geçerlik ve güvenirlik çalışmaları. [Developing the Social Media Addiction Scale: Validity and Reliability Studies] *Addicta: The Turkish Journal on Addictions*, *5*, 507–525. http://dx.doi.org/10.15805/addicta.2018.5.3.0046 boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer - Mediated Communication, 13*, 2010 - 2030. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Brunborg, G. S., Mentzoni, R. A., Molde, H., Myrseth, H., Skouverøe, K. J. M., Bjorvatn, B., & Pallesen, S. (2011). The relationship between media use in the bedroom, sleep habits, and symptoms of insomnia. *Journal of Sleep Research, 20*, 569-575. https://doi.org/10.1111/j.1365-2869.2011.00913.x

Cakir Balta, O. & Horzum M. B. (2008). İnternet bağımlılığı testi. [Internet addiction test]. *Eğitim Bilimleri ve Uygulama, 7*(13), 87-102

Can, L., & Kaya, N. (2016). Social networking sites addiction and the effect of attitude towards social network advertising. *Procedia-Social and Behavioral Sciences*, *235*, 484-492. https://doi.org/10.1016/j.sbspro.2016.11.059

Cao, F., & Su, L., (2006). Internet addiction among Chinese adolescents: prevalence and psychological features. *Child Care Health and Development. 33*(3), 275–281. https://doi.org/10.1111/j.1365-2214.2006.00715.x

Cecilia, C., & Yee-lam L. A., (2014). Internet addiction prevalence and quality of (real) life: a meta-analysis of 31 nations across seven world regions, *Cyberpsychology, Behavior, and Social Networking. 17*(12), 755-760.

Cheung, L. M., & Wong, W. S. (2011). The effects of insomnia and internet addiction on depression in Hong Kong Chinese adolescents: An exploratory cross-sectional analysis. *Journal of Sleep Research, 20*, 311–317. http://dx.doi.org/10.1111/j.1365-2869.2010.00883.x

Choi, S. B., & Lim, M. S. (2016). Effects of social and technology overload on psychological well-being in young South Korean adults: The mediatory role ofsocial network service addiction. *Computers in Human Behavior, 61*, 245-254.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal Bilimler için Çok Değişkenli İstatistik SPSS ve LISREL Uygulamaları*. Ankara: Pegem Akademi.

De Beer, M. (2004). Use of differential item functioning (DIF) analysis for bias analysis intest construction. *South African Journal of Industrial Psychology, 30*(4), 52-58.

Dewald, J. F., Meijer, A. M., Oort, F. J., Kerkhof, G. A., & Bögels, S. M. (2010) The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: a meta-analytic review. *Sleep Medicine Reviews, 14*, 179-189.

Echeburúa, E., & De, Corral. P. (2010). Addiction to new technologies and to online socialnetworking in young people: A new challenge. *Adicciones,22*(2), 91-95.

Erkuş, A. (2012). *Psikolojide Ölçme ve Ölçek Geliştirme-1 temel kavramlar ve işlemler*. Ankara: Pegem Akademi.

Esgi, N. (2016). Development of social media addiction test (SMART17). *Journal of education and training studies*. *4*(10), 174-181

Gao, W., Liu, Z., & Li, J. (2017). How does social presence influence SNS addiction? A belongingness theory perspective. *Computers in Human Behavior 77* (2017) 347-355. http://dx.doi.org/10.1016/j.chb.2017.09.002

Griffiths, M. D. (2013). Social networking addiction: emerging themes and issues. *Journal of Addiction Research & Therapy*, *4*(5). http://dx.doi.org/10.4172/2155-6105.1000e118

Griffiths, M. D., Kuss, D. J., & Demetrovics, Z. (2014). Social networking addiction: anoverview of preliminary findings. *In Behavioral addictions. Criteria, evidence, and treatment* (pp. 119-141). New York: Elsevier.

Gültaş, M. (2014). Work Discipline Compound Personality Scale Development with Item Response Theory. Unpublished dissertation thesis. Middle East Technical University, Graduate School of Social Sciences, Ankara.

Huang, H., & Leung, L. (2009). Instant messaging addiction among teenagers in China: Shyness, alienation, and academic performance decrement. *Cyberpsychology and Behavior,12*(6), 675-679. http://dx.doi.org/10.1089/cpb.2009.0060

Huang, X., Zhang, H., Li, M., Wang, J., Zhang, Y., & Tao, R., (2010). Mental health, personality, and parental rearing styles of adolescents with internet addiction disorder. *Cyberpsychology, Behavior, and Social Networking. 13*(4), 401-406. DOI: 10.1089=cyber.2009.0222

Kang, I., Shin, M. M., & Park, C. (2013). Internet addiction as a manageable resource:A focus on social network services. *Online Information Review, 37*(1), 28-41.

Kim, E. J., Namkoong, K., Ku, T., & Kim, S.J., (2008). The relationship between online game addiction and aggression, self-control and narcissistic personalitytraits. *European Pschiatry, 23*, 212-218.

Kim, K., Ryu, E., Chon, M. Y., Yeun, E. J., Choi, S. Y., Seo, J. S., &Nam, B.W., (2006). Internet addiction in Korean adolescents and its relation to depression andsuicidal ideation: a questionnaire survey. *Intertional Journal of Nursing Studies, 43*, 185-192.

Kuss, D. J., & Griffiths, M. D. (2011). Online social networking and addiction-A review of the psychological literature. *International Journal of Environmental Research and Public Health, 8,* 3528-3552. DOI:10.3390/ijerph8093528

Kuss, D. J., & Griffiths, M. D. (2012). Internet gaming addiction: A systematic reviewof empirical research. *International Journal of Mental Health and Addiction, 10*(2), 278-296.

Kuşay, Y. (2013). *Sosyal Medya Ortamında Çekicilik ve Bağımlılık- Facebook Üzerine Bir Araştırma*. İstanbul: Beta Yayıncılık.

LaRose, R., Connolly, R., Lee, H., Li, K., & Hales, K. D. (2014). Connection overload? Across cultural study of the consequences of social media connection. *Information Systems Management*, *31*(1), *59*-73. https://doi.org/10.1080/10580530.2014.854097

Lavin, M. J., Yuen, C.N., Weinman, M., & Kozak, K., (2004). Internet dependence in the collegiate population: The role of shyness. *CyberPsychology & Behavior. 7*, 379-383.

Pantic, I. (2014). Online social networking and mental health. *Cyberpsychology, Behavior, and Social Networking, 17*(10), 652-657. DOI: 10.1089/cyber.2014.0070

Park, S. K., Kim, J. Y., & Cho, C. B. (2008). Prevalence of Internet addiction and correlations with family factors among South Korean adolescents. *Adolescence,43*(172), 895-909.

Ryan, T., Chester, A., Reece, J., & Xenos, S. (2014). The uses and abuses of Facebook: A review of Facebook addiction. *Journal of Behavioral Addictions 3*(3), 133-148. DOI: 10.1556/JBA.3.2014.016

Salehan, M., & Negahban, A. (2013). Social networking on smartphones: Whenmobile phones become addictive. *Computers in Human Behavior, 29*(6), 2632-2639.

Satici, S. A., & Uysal, R. (2015). Well-being and problematic Facebook use. *Computers in Human Behavior*, *49*, 185-190. DOI:10.1016/j.chb.2015.03.005

Suganuma, N., Kikuchi, T., Yanagi, K., Yamamura, S., Morishima, H., Adachi, H., Kumanogo, T., Mikami, A., Sugita, Y., & Takeda, M. (2007). Using electronic media before sleep can curtail sleep time and result in self-perceived insufficient sleep. *Sleep and Biological Rhythms, 5*, 204-214.

Şahin, C. (2018). Social media addiction scale - Student form: The reliability and validity study, TOJET: *The Turkish Online Journal of Educational Technology,17*(1), 169-182. ERIC Number: EJ1165731

Şahin, C., & Yağcı, M. (2017). Sosyal medya bağımlılığı ölçeği- Yetişkin formu: Geçerlilik ve güvenirlik çalışması [Social Media Addiction Scale - Adult Form: The Reliability and Validity Study Social Media Addiction Scale - Adult Form: The Reliability and Validity Study], *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi (KEFAD), 18*(1), 523-538.

Taş, İ. (2017). Ergenler için sosyal medya bağımlılığı ölçeği kısa formunun (SMBÖ-KF) geçerlik ve güvenirlik çalışması, [The Study of Validity And Reliability of The Social Media Addiction Scale Short Form For Adolescents], *Online Journal of Technology Addiction & Cyberbullying*, *4*(1), 27-40.

Tekin, C., Gunes, G., & Colak, C. (2014). Adaptation of problematic mobile phone use scale to Turkish: a validity and reliability study. *Medicine Science*, *3*(3), 1361-81. DOI: 10.5455/medscience.2014.03.8138

Turel, O., & Serenko, A. (2012). The benefits and dangers of enjoyment with socialnetworking websites. *European Journal of Information Systems*, *21*(5), 512-528.

Tutgun-Ünal, A. (2015). *Sosyal medya bağımlılığı: üniversite öğrencileri üzerine bir araştırma*. [Social media addiction: a research on university Students], (Yayınlanmamış Doktora Tezi) Marmara Üniversitesi Sosyal Bilimler Enstitüsü, Gazetecilik Ana Bilim Dalı, Bilişim Bilim Dalı. İstanbul.

Van den Eijnden, R. J. J. M., Lemmens, J. S., & Valkenburg, P. M. (2016). The social media disorder scale. *Computers in Human Behavior*, *61*, 478-487.

Vilca, L. W., & Vallejos, M. (2015). Construction of the risk of addiction to socialnetworks scale (Cr. ARS). *Computers in Human Behavior*, *48*, 190-198.

We are Social (2017). Digital in 2017 Global Overview Report. https://wearesocial.com/uk/special-reports/digital-in-2017-global-overview

Wu, A. D., Li, Z. ve Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, *12*, 1-26.

Yang, S., Liu, Y., & Wei, J. (2016). Social capital on mobile SNS addiction: A perspective from online and offline channel integrations. *Internet Research*, *26*(4), 982-1000.

Young, K. S. (1998). Internet addiction: The emergence of a new clinical disorder. *Cyberpsychology and Behavior*, *1*(3), 237-244. https://doi.org/10.1089/cpb.1998.1.237

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 415–435
https://dx.doi.org/10.21449/ijate.539487

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

Efficiency Measurement with Network DEA: An Application to Sustainable Development Goals 4

Deniz Koçak [1,*], **Hasan Türe** [1], **Murat Atan** [1]

[1] Department of Econometrics, Ankara Hacı Bayram Veli University, Ankara, Turkey

ARTICLE HISTORY

Received: 13 March 2019

Revised: 23 July 2019

Accepted: 18 August 2019

KEYWORDS

Relational network DEA,

Sustainable development goals 4,

Educational economy

Abstract: Education is the core of the factors that improved people for a better lifestyle and increases the level of society' development. Quality education is one of the most vital goals of Sustainable Development Goals (SDGs) due to actualizing these factors. Using relational network data envelopment analysis (DEA), which have three interrelated substages, this current paper computes the educational economy efficiency of the Organisation for Economic Co-operation and Development (OECD) countries bearing in mind the characteristics related to SDGs. The contribution of our study is the use of a novel approach to computing the educational economy efficiency using relational network DEA with GAMS. Even though some interesting differences reveal in the efficiency of the countries, the findings show that countries with high-efficiency scores are clustered around countries like Latvia, Slovenia, and Korea.

1. INTRODUCTION

Performance evaluation is a crucial phenomenon for countries with regard to determining the current situation and finding an efficient process that ameliorated this situation. The good performance of countries on social, economic and health issues is possible through the acquisition of a quality education that influences directly the lives and sustainable development of human. Well-educated human capital can be considered of the engine of the production process for new discoveries, ideas, development and eventually new value-added productions.

In recent years, numerous studies have been examined as a result of the widespread interest in education. The United Nations emphasized that education for all is always an inseparable part of the agendas of both Millennium Development Goals (MDGs) and Sustainable Development Goals (SDGs) (United Nations, 2015). MDGs are expired at the end of 2015. SDGs is a new agenda integrated into MDGs and covering a 15 years period for the post-2015 with 17 goals and 169 interrelated targets in global developments efforts in social, economic and environmental areas (Griggs et al., 2013; Le Blanc, 2015). Quality Education, which is defined as *Ensure Inclusive and Equitable Quality Education and Promote Lifelong Learning Opportunities for All*, is the fourth goal of SDGs. Besides, SDGs have 10 targets comprising many different features of education, which aimed at educating the people for enhancing their

CONTACT: Deniz Koçak ✉ deniz.kocak@hbv.edu.tr Department of Econometrics, Ankara Hacı Bayram Veli University, Ankara, Turkey.

ISSN-e: 2148-7456 /© IJATE 2019

individual well-being and socio-economic status. By 2030, these agendas pave the way of the creation of the societies with strong sustainable education and culture thanks to the effective and functional learning outcomes of these objectives (Hopkins & McKeown, 2002; Sterling, 2001).

Quality education is a cornerstone that ensures the human's sustainable development. Although the global awareness is existed for the importance of education, more than 265 million children are out of school and 22% of them are of primary school age and roughly the same number of them as will be out of school (UN, 2018; UNICEF, 2016). As a percentage of GDP, Latvia spends more on primary education than any other Organisation for Economic Co-operation and Development (OECD) country, followed closely by Slovenia and Poland. However, considering the education expenditure to secondary and tertiary education, Denmark takes the first place. Turkey spend the least as a percentage of their GDP on primary and secondary education. The share of public expenditure per student on tertiary education in Korea is also among the lowest, especially at the tertiary level. Estonia located one of the top performers in Programme for International Student Assessment (PISA) followed by Finland and Korea. In the light of these findings, the analysis of the educational economy efficiency of the OECD countries is crucial from the point of view of policymakers, officials and researchers who are concerned with education in both regional or worldwide (Abbott & Doucouliagos, 2003; Barra, Lagravinese & Zotti 2018; Worthington, 2001).

While evaluating the educational economy performance of the countries, any study which consider the linkages between substages of education could not be encountered. Besides, the lack of the efficient analysis or metrics in measuring the performance of the countries makes it difficult to carry out the comparisons of the countries. For intend to fill this gap, the efficiency of most OECD countries is measured towards to indicators based on education economics, employment and PISA (Programme for International Student Assessment) data. This measurement gives critical feedback to international studies whether the education system is quality and well design. Besides, an important focus of this study is that a newest developed theory (relational network DEA) is used for this intent. The relational network DEA can assess countries' education performance from a multistage efficiency and effectiveness perspective and further examine their education performance from a multidimensional viewpoint. Unlike previous studies, we not only directly investigate the relationship between inputs and outputs, but also consider the linkages between education substages (primary, secondary, and tertiary) by means of this methodology. Measuring the efficiency at these disaggregated levels is of great importance as it reflects realistically the concept of that education materializes at the student level (Ruggiero, 2006).

The structure of this paper is organized as follows: Section 2 discussed some relevant indicators of the educational economy. Section 3 describes the methodology. Section 4 presents the research findings obtained from running the GAMS codes. The final section presents a brief summary of the results.

1.1. Literature Review

Education efficiency assessment has become a core research to understand progress difficulties owing to the education services supported by government almost every country. To this end, different methodological approaches have showed up over the past few decades. Although the efficiency measurement techniques have been applied to many different types of institutions, but the studies on an international framework with whole countries as units of observation are rarely encountered (Afonso & Aubyn, 2006). These studies have generally concentrated on how to assign educational resource inputs to improve output performance efficiently. Moreover, it is well known that the education expenditure as input (see Afonso et al., 2005; Aubyn, 2003; Ciro & Garcia, 2018; Gupta & Verhoeven, 2001; Hanushek & Kimko, 2000; Lee & Barro, 2001), PISA score as output (see Afonso & Aubyn, 2005; Afonso & Aubyn, 2006; Aristovnik, 2012; Jafarov & Gunnarsson, 2008), and employment rate as output (see Afonso & Aubyn, 2006; Chen & Wu, 2007; Lavrinovicha et al., 2015) are the important factors of measurement of educational economy efficiency.

Generally, two type of decision making units (DMUs) have been used to assess the level of efficiency with respect to government expenditure on education in these studies. In the first group, the micro education level (university, school etc.) consider as DMUs. Furthermore, the macro level approaches in which countries are selected as DMUs are included in the second group.

There are various studies for assessing the education efficiency at micro level. Ramzi, Afonso and Ayadi (2016) used DEA for reveal the relationship between school resources and student performance. It is find that inefficiency in education was associated with the poverty in the governorates. Kashim et al. (2017) measured the efficiency of a university faculty in Malaysia by using a network DEA model. They selected several inputs including number of academicians (professors, associate professors, senior lecturers, lecturers, foreign academic staff, non-academic staff) and expenses. The outputs included number of graduates (from undergraduate program, master program, and Ph.D. program), publications, grants, main researchers based on different types of grants, expert lecturers, collaboration activities done under MoU/LoI. Qin and Du (2018) applied the network DEA approach to assess the effectiveness of the universities' research and development (R&D) performance.

Yang et al. (2018) investigated the inefficiency and productivity of Chinese universities, using two-stage network process over the period of 2010-2013. They used R&D funds, teaching and research staff, and government block funds as input and number of SCI/SSCI publications, the total number of students, patents, and the other intellectual property forms as output in the first stage. In the second stage, the number of patents and other intellectual properties, which is already used as output in the previous stage, and the staff of the application of R&D outputs and technology services were used as input; total income was used as output.

When viewed from macro level approaches, Afonso and Aubyn (2006) examined the efficiency of expenditure in education for the 25 mostly OECD countries by using a semiparametric model of a two-step DEA/Tobit analysis. PISA scores, education spending per student, number of teachers, and time spent at school were used as input. These indicators are similar to Sylwester (2002) which was revealed the government spends on education encourage income equality and Wasylenko and McGuire (1985) that the government expenditure on education increase the employment rate.

Guironnet and Peypoch (2018) seek an answer how institutional factors affect the productivity of university by using hierarchical DEA following distinctions: urban/rural areas and public/private universities. Ciro and Garcia (2018) emphasized that most discussions have concentrate on the importance of increasing public expenditure on education covers 37 countries. They measured the efficiency of public secondary education expenditure using a two-step semi-parametric DEA methodology. Private spending (%GDP), and government expenditures (%GDP per capita) were selected as input in the first model. Furthermore, the enrolment rates and PISA scores were used as outputs; the teacher-pupil ratio was used as input in second model.

It is important to note that PISA scores of the countries are a remarkable indicator in connection with the test is internationally validity. In this context, Aristovnik (2012) used the average data for 1999-2007 period to show that the long-term efficiency measures as the effects of ICT (Information and Communication Technology) are characterized by time lags. The study find that ICT had a significant impact on education sector in the selected EU-27 and OECD countries.

2. METHOD

2.1. Relational Network DEA Model

The relational network DEA model accounts for both the efficiency of a system and the system's interrelated substages. Thus, the drawbacks of the traditional DEA models that neglects of interrelated substages can be eliminated. Besides, the overall steps of the traditional DEA models that are so-called "black box" can been made explicit.

This study uses the relational network DEA model is comprised of a series of three substages under the assumption of the constant return to scale and output-oriented. The fact that the aim is to increase output rather than input reduction in the educational economy efficiency reveals that the output-oriented model is the appropriate tool (Johnes, 2006).

Figure 1 presents the relational network DEA structure with inputs $X_i, i = 1, 2, \dots, m$, intermediate products $Z_p, p = 1, 2, \dots, q$ and $T_l, l = 1, 2, \dots, d$, and outputs $Y_r, r = 1, 2, \dots, s$.

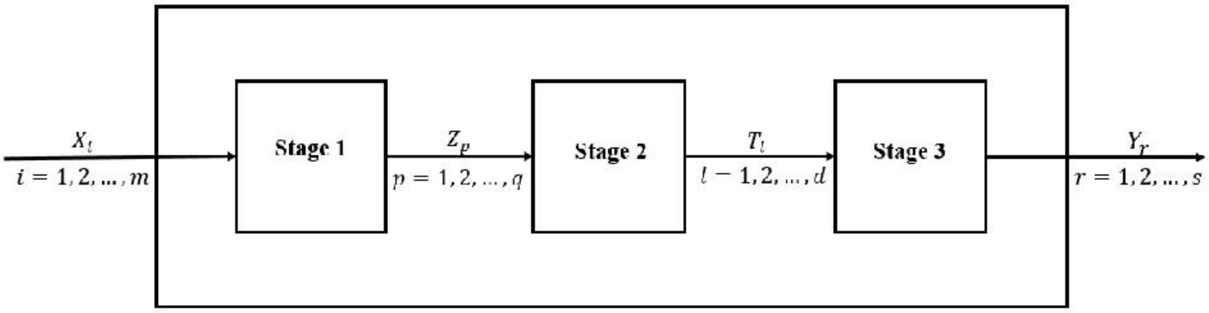


Figure 1. Relational network DEA structure

We define X_i and Y_r the *i*th input and *r*th output of the *j*th DMU by denoting the *i*, *r*, *j* indexes of input, output and DMU. The intermediate products 1, which is the outputs of the first stage and the inputs of the second stage, and the intermediate products 2, which is the outputs of the second stage and the inputs of the third stage, are represented by respectively Z_p and T_l by denoting the *p*, *l*, *j* indexes of intermediate products 1, intermediate products 2 and DMU.

The substages efficiency calculated by E_k^1, E_k^2 and E_k^3, and the overall efficiency of the system can be calculated by $E_k = E_k^1 \times E_k^2 \times E_k^3$. The linear program model of the overall efficiency and its constraint proposed by Kao (2009) is:

$$E_k = \max_{u_r, w_p, \gamma_l, v_i} \sum_r u_r Y_r$$

Subject to:

$$\begin{aligned}
&\sum_i v_i X_{i_i} = 1 \ \forall k,\\
&\sum_r u_r Y_r - \sum_i v_i X_i \le 0 \ \forall j,\\
&\sum_l \gamma_l T_l - \sum_i v_i X_i \le 0 \ \forall j,\\
&\sum_p w_p Z_p - \sum_i v_i X_i \le 0 \ \forall j,\\
&\sum_r u_r Y_r - \sum_p w_p Z_p \le 0 \ \forall j,
\end{aligned} \qquad (1)$$

$$\sum_l \gamma_l T_l \; - \sum_p w_p Z_p \;\leq 0 \;\; \forall j,$$
$$u_r, w_p, \gamma_l, v_i \geq 0 \;\; \forall r, p, l, i.$$

The aim of the optimal multipliers u_r, w_p, γ_l, v_i is unique, in the first instance the overall efficiency namely Model (1) is calculated. Then the efficiencies of the substages must be calculated. In this study, after the measurement of the overall efficiency, the second and third stage will be calculated. With the help of the $E_k = E_k^1 \times E_k^2 \times E_k^3$, the efficiency of the first stage can be obtained as $E_k^1 = E_k/(E_k^2 \times E_k^3)$. Model (2) shows the linear program of the efficiency of the third stage and its constraint:

$$E_k^3 = \max_{u_r, w_p, \gamma_l, v_i} \sum_{r=1}^{s} u_r Y_r$$

Subject to:

$$\begin{aligned}
&\sum_{l=1}^{t} \gamma_l H_j = 1,\\
&\sum_r u_r Y_r \; - E_k \sum_i v_i X_i \; = 0 \; \forall j,\\
&\sum_r u_r Y_r \; - \sum_i v_i X_{ij} \leq 0 \;\; \forall j,\\
&\sum_l \gamma_l T_l \; - \sum_i v_i X_i \;\; \leq 0 \;\; \forall j,\\
&\sum_p w_p Z_p \; - \sum_i v_i X_i \;\; \leq 0 \;\; \forall j,\\
&\sum_r u_r Y_r \; - \sum_p w_p Z_p \;\; \leq 0 \;\; \forall j,\\
&\sum_l \gamma_l T_l \; - \sum_p w_p Z_p \;\; \leq 0 \;\; \forall j,\\
&u_r, w_p, \gamma_l, v_i \geq 0 \;\; \forall r, p, l, i.
\end{aligned} \tag{2}$$

If the first constraint and the objective function of Model (2) is expressed as $\sum_{p=1}^{q} w_p Z_p \;\; = 1$ and $\sum_{l=1}^{t} \gamma_l T_l$, the efficiency of the second stage can be obtained. Otherwise, if the first constraint and the objective function of Model (2) is expressed as $\sum_{i=1}^{m} v_i X_i \;\; = 1$ and $\sum_{l=1}^{t} w_p Z_p$, the efficiency of the first stage can be obtained.

2.2. Data

In this study, the data that express the sub-objectives of the SDG 4 and can be used to measure the educational economy efficiency of the OECD countries are taken into consideration. The 30 OECD countries are the DMUs in the analysis. Besides, the inputs, intermediate products and outputs of the network structure are expressed as Table 1:

Table 1. The inputs, intermediate products and outputs[†].

X_1:	Government expenditure per primary student (% of GDP per capita)	2013-14
Z_1:	PISA science performance (mean)	2015
Z_2:	Government expenditure per secondary student (% of GDP per capita)	2013-14
T_1:	Employment rate for upper secondary level (% of 25-64 year-olds)	2013-15
T_2:	Government expenditure per tertiary student (% of GDP per capita)	2013-14
Y_1:	Employment rate for tertiary level (% of 25-64 year-olds)	2013-15

These indicators based on the levels defined by International Standard Classification of Education (ISCED) are taken are will be used in the substages of the network structure. The

[†] The data that cover 2013-2015 period were collected from the database of OECD.

reason is that these standard international education levels provide unity for measuring the performance of the students (Johnes et al., 2017). PISA are used in the analysis that is the reason why the quality of education can be measured by the achievement of students via the scores on international test represented the cross-country variations in cognitive skills of the students and thereby the differences in the quality of the future labour force (Lee & Barro, 2001). Besides, the government expenditure that placed in primary, secondary and tertiary levels is selected in relation to the country's sources of educational finance (Riddell, 1993). Figure 2 presents the framework of the network structure modeled as a three-stage process.

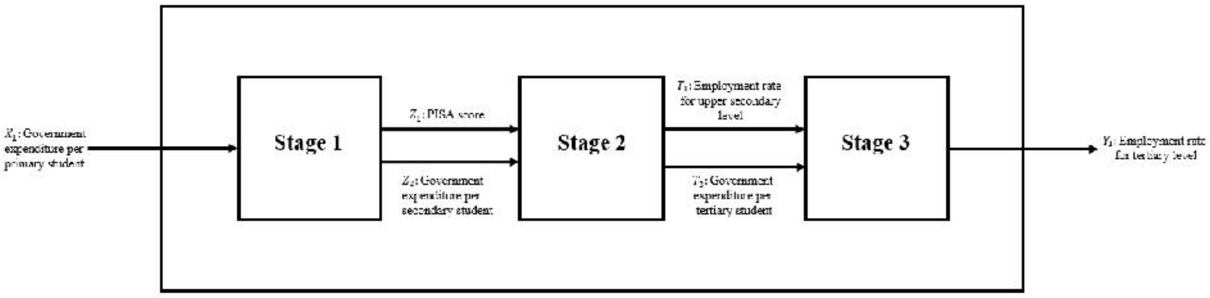

X_1: Government expenditure per primary student
Stage 1
Z_1: PISA score
Z_2: Government expenditure per secondary student
Stage 2
T_1: Employment rate for upper secondary level
T_2: Government expenditure per tertiary student
Stage 3
Y_1: Employment rate for tertiary level

Figure 2. Relational network DEA structure for SDG 4

Table 2. Data

DMU	Country	X_1	Z_1	Z_2	T_1	T_2	Y_1
1	Australia	18.5932	509.9939	16.7917	77.6700	22.5295	83.1767
2	Austria	23.4497	495.0375	27.3519	75.9433	36.1737	85.4733
3	Belgium	22.4726	501.9997	25.7987	72.8533	33.0074	84.4900
4	Chile	15.0754	446.9561	15.1517	71.7000	17.3957	84.1950
5	Czech Republic	15.5312	492.8300	23.5477	77.7067	21.6123	84.7400
6	Denmark	25.6117	501.9369	28.2276	79.6433	44.6629	86.2533
7	Estonia	21.4939	534.1937	20.1925	75.7733	27.7667	84.2533
8	Finland	21.0185	530.6612	27.2000	73.2000	35.4852	83.3400
9	France	18.0165	494.9776	26.8059	72.8867	35.0586	84.0300
10	Germany	17.9128	509.1406	23.4938	79.4667	37.5885	87.9900
11	Hungary	18.4491	476.7475	19.5469	71.5033	23.8811	81.6667
12	Iceland	24.4524	473.2301	18.3453	87.0433	25.6424	91.2133
13	Ireland	16.7971	502.5751	21.6000	67.6133	25.2509	81.1100
14	Israel	21.4471	466.5528	16.9659	72.4467	19.4927	85.9100
15	Italy	21.3017	480.5468	23.0962	69.8567	26.1989	78.1000
16	Korea	23.9656	515.8099	23.3231	71.9867	13.7440	77.4733
17	Latvia	31.2500	490.2250	29.6703	70.9233	22.9545	85.1000
18	Mexico	14.8592	415.7099	16.4109	70.6100	40.4662	80.3367
19	New Zealand	18.6102	513.3035	22.3160	80.9400	27.9934	86.8167
20	Norway	19.9855	498.4811	24.3500	81.2633	38.0238	89.5500
21	Poland	26.4191	501.4353	21.7531	66.1867	24.8147	86.0367
22	Portugal	23.6089	501.1001	15.1691	77.4400	25.4516	82.2067
23	Slovak Republic	19.4372	460.7749	18.7881	71.2133	20.7731	79.9067
24	Slovenia	28.8696	512.8636	25.5421	69.5467	21.1539	83.8000
25	Spain	17.7485	492.7861	22.4697	66.0233	22.6820	77.3800
26	Sweden	25.3405	493.4224	24.6998	84.1767	43.4855	89.1367
27	Switzerland	25.3405	505.5058	25.4500	81.2233	38.1637	88.1067
28	Turkey	13.3391	425.4895	14.7689	61.8800	24.2958	76.4267
29	United Kingdom	22.8298	509.2215	22.6604	79.5733	37.0920	84.9200
30	United States	19.8534	496.2424	22.5901	68.0967	24.6532	80.5533

In stage 1, government expenditure for primary level and PISA scores are taken into consideration. In stage 2 and stage 3, government expenditures, employment rates respectively for upper secondary and tertiary levels are taken into consideration. In this case, we can measure the efficiency of stage 1 of each OECD country among the set of DMUs using X_1 as input and Z_1, Z_2 as outputs. The efficiencies of the stage 2 can be measured using Z_1, Z_2 as input and T_1, T_2 as outputs. Similarly, the efficiencies of the stage 3 can be measured using T_1, T_2 as input and Y_1 as outputs. The overall efficiency is also measured using X_1 as input and Y_1 as outputs with Model (1). Table 2 presents the network structure for the implementation of the SDG 4 and the data.

3. RESULT and FINDINGS

As mentioned in the literature section, a number of studies have shown that there is a positive link between government expenditure on education and employment. This study reveals that the network DEA model can be used with the aim of measuring the efficiency of the OECD countries from an educational economy perspective. The overall and substages efficiencies are calculated for each country with the Model (1) and Model (2) using the GAMS code in Appendix A and Appendix B.

After running the GAMS codes, the efficiency scores and the rank of countries are shown in Table 3 and Figure 3. The ranking at the overall efficiency scores (E_k) shows that Latvia, Slovenia and Korea at the top-three countries in terms of network structure's indicators. Additionally, Turkey, Chile and Czechia are found as the lowest three countries. Broadly speaking, the findings are verified that the developed countries in data set are carried to an upper order in the ranking. Conversely, the developing countries such as Turkey, Chile and Mexico are located at the lower in the ranking. Amazingly, Korea is located at third place notwithstanding the country is developing.

Table 3 shows that Sweden, Iceland, United Kingdom and Portugal are at the highest order of ranking in stage-1. At first look the rank of Portugal are demonstrating encouraging results. But, Portugal is one of the countries that has made the fastest progress in improving educational attainment such as PISA scores (OECD, 2012). In this context, there is no doubt that an increase in PISA scores and government expenditure for related level will lead to an enhancement in the efficiency of stage-1 due to the impact of the output-oriented model in the analysis.

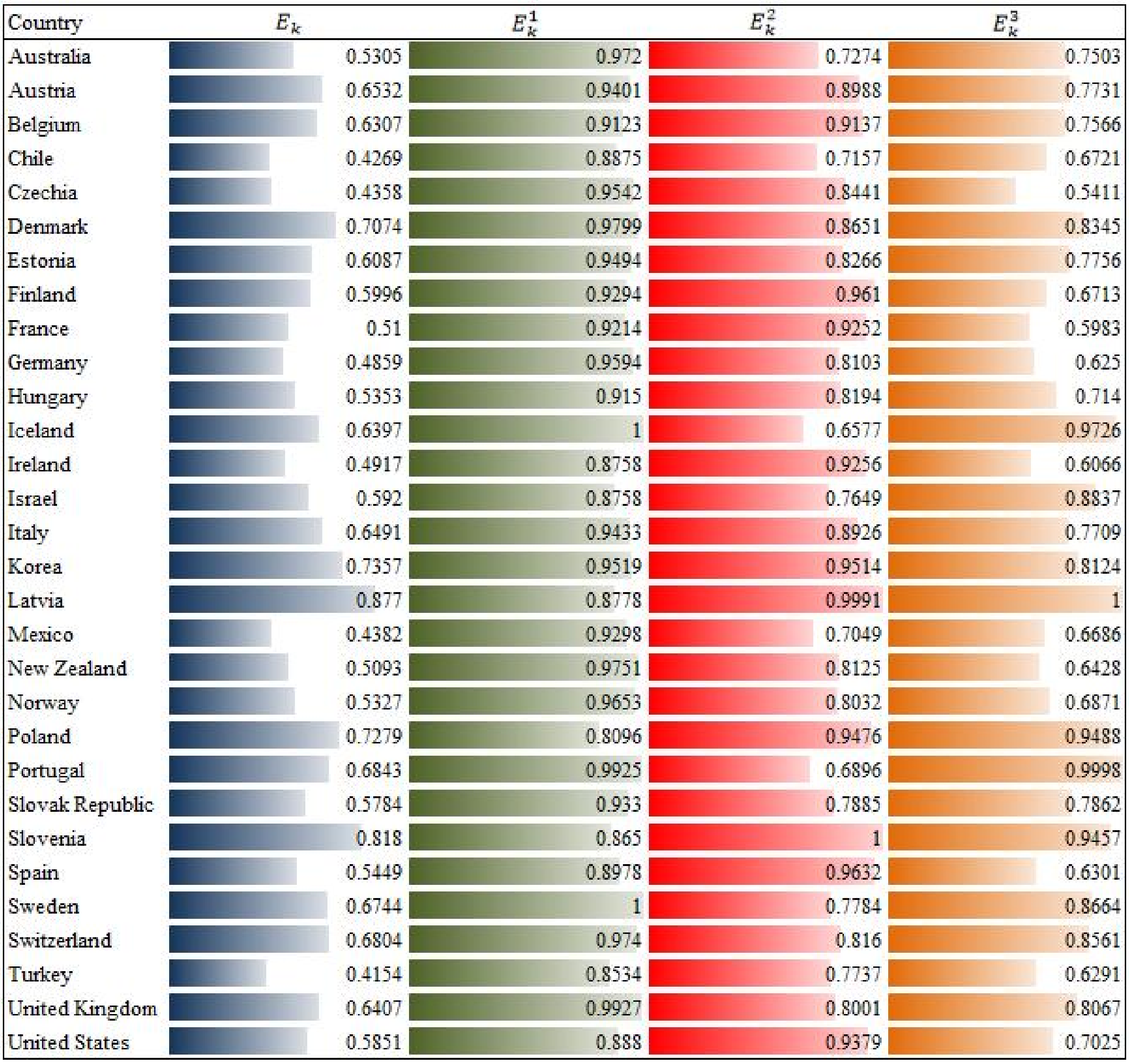
In stage-2, Slovenia has been found an efficient country with regard to indicators of the educational economy. Besides, Latvia, and Spain have an efficiency score that is very close to 1.0000. However, Iceland, Portugal, and Mexico have the lowest efficiency scores. Considering the structure of Figure 3, high investments in education accelerates the growth of countries, and the growth of the country maintains the employment rates (Domar, 1946; Landau, 1983).

In stage-3, Latvia has the highest efficiency score ($E_k = 1.0000$). Following this country, it is seen that Portugal and Iceland have respectively 0.9998 and 0.9726 efficiency score. The fact that those countries are in the top three can be owing to having high employment rate and budgeting high government expenditure per student. On the other hand, Czechia, France, and Ireland have respectively 0.5411, 0.5983 and 0.6066 efficiency scores. This indicates that these countries need improvement in the indicators that used for the network structure model.

Table 3. *Efficiency scores*[*]

DMU	Country	E_k	E_k^1	E_k^2	E_k^3
1	Australia	0.5305 (22)	0.9720 (7)	0.7274 (26)	0.7503 (17)
2	Austria	0.6532 (9)	0.9401 (14)	0.8988 (11)	0.7731 (14)
3	Belgium	0.6307 (13)	0.9123 (20)	0.9137 (10)	0.7566 (16)
4	Chile	0.4269 (29)	0.8875 (23)	0.7157 (27)	0.6721 (21)
5	Czechia	0.4358 (28)	0.9542 (10)	0.8441 (14)	0.5411 (30)
6	Denmark	0.7074 (5)	0.9799 (4)	0.8651 (13)	0.8345 (9)
7	Estonia	0.6087 (14)	0.9494 (12)	0.8266 (15)	0.7756 (13)
8	Finland	0.5996 (15)	0.9294 (17)	0.9610 (4)	0.6713 (22)
9	France	0.5100 (23)	0.9214 (18)	0.9252 (9)	0.5983 (29)
10	Germany	0.4859 (26)	0.9594 (9)	0.8103 (19)	0.6250 (27)
11	Hungary	0.5353 (20)	0.9150 (19)	0.8194 (16)	0.7140 (18)
12	Iceland	0.6397 (12)	1.0000 (1)	0.6577 (30)	0.9726 (3)
13	Ireland	0.4917 (25)	0.8758 (25)	0.9256 (8)	0.6066 (28)
14	Israel	0.5920 (16)	0.8758 (26)	0.7649 (25)	0.8837 (6)
15	Italy	0.6491 (10)	0.9433 (13)	0.8926 (12)	0.7709 (15)
16	Korea	0.7357 (3)	0.9519 (11)	0.9514 (5)	0.8124 (10)
17	Latvia	0.8770 (1)	0.8778 (24)	0.9991 (2)	1.0000 (1)
18	Mexico	0.4382 (27)	0.9298 (16)	0.7049 (28)	0.6686 (23)
19	New Zealand	0.5093 (24)	0.9751 (5)	0.8125 (18)	0.6428 (24)
20	Norway	0.5327 (21)	0.9653 (8)	0.8032 (20)	0.6871 (20)
21	Poland	0.7279 (4)	0.8096 (29)	0.9476 (6)	0.9488 (4)
22	Portugal	0.6843 (6)	0.9925 (3)	0.6896 (29)	0.9998 (2)
23	Slovak Republic	0.5784 (18)	0.9330 (15)	0.7885 (22)	0.7862 (12)
24	Slovenia	0.8180 (2)	0.8650 (27)	1.0000 (1)	0.9457 (5)
25	Spain	0.5449 (19)	0.8978 (21)	0.9632 (3)	0.6301 (25)
26	Sweden	0.6744 (8)	1.0000 (1)	0.7784 (23)	0.8664 (7)
27	Switzerland	0.6804 (7)	0.9740 (6)	0.8160 (17)	0.8561 (8)
28	Turkey	0.4154 (30)	0.8534 (28)	0.7737 (24)	0.6291 (26)
29	United Kingdom	0.6407 (11)	0.9927 (2)	0.8001 (21)	0.8067 (11)
30	United States	0.5851 (17)	0.8880 (22)	0.9379 (7)	0.7025 (19)

[*] The values in parentheses are rank values of the countries.



Country	E_k	E_k^1	E_k^2	E_k^3
Australia	0.5305	0.972	0.7274	0.7503
Austria	0.6532	0.9401	0.8988	0.7731
Belgium	0.6307	0.9123	0.9137	0.7566
Chile	0.4269	0.8875	0.7157	0.6721
Czechia	0.4358	0.9542	0.8441	0.5411
Denmark	0.7074	0.9799	0.8651	0.8345
Estonia	0.6087	0.9494	0.8266	0.7756
Finland	0.5996	0.9294	0.961	0.6713
France	0.51	0.9214	0.9252	0.5983
Germany	0.4859	0.9594	0.8103	0.625
Hungary	0.5353	0.915	0.8194	0.714
Iceland	0.6397	1	0.6577	0.9726
Ireland	0.4917	0.8758	0.9256	0.6066
Israel	0.592	0.8758	0.7649	0.8837
Italy	0.6491	0.9433	0.8926	0.7709
Korea	0.7357	0.9519	0.9514	0.8124
Latvia	0.877	0.8778	0.9991	1
Mexico	0.4382	0.9298	0.7049	0.6686
New Zealand	0.5093	0.9751	0.8125	0.6428
Norway	0.5327	0.9653	0.8032	0.6871
Poland	0.7279	0.8096	0.9476	0.9488
Portugal	0.6843	0.9925	0.6896	0.9998
Slovak Republic	0.5784	0.933	0.7885	0.7862
Slovenia	0.818	0.865	1	0.9457
Spain	0.5449	0.8978	0.9632	0.6301
Sweden	0.6744	1	0.7784	0.8664
Switzerland	0.6804	0.974	0.816	0.8561
Turkey	0.4154	0.8534	0.7737	0.6291
United Kingdom	0.6407	0.9927	0.8001	0.8067
United States	0.5851	0.888	0.9379	0.7025

Figure 3. Efficiency scores of the countries

To sum up, the reason for the high ranks of the countries is that the employment rate can be increased by the awareness of the quality education. Within this framework, these countries can enhance their efficiency score by designing educational economy policies toward strengthening the employment rate per each stage.

4. DISCUSSION and CONCLUSION

The government expenditure on education can be regarded as one of the most important indicators influence on increasing employment growth. On a priori grounds, it is not always possible to measure the efficiency of the countries on how government expenditure affects the employment rate. However, it is noteworthy that the evaluation of the countries as a whole in terms of educational economy for academic literature and policy-making studies. Besides, indicators such as government expenditures on education and employment rates at education levels determined by the OECD can be used directly in the evaluation of the educational performance of countries. However, these indicators used alone are not sufficient to determine the educational economy performance of a country. In this context, situations related to the different economic, social and cultural conditions of the nations at the micro level should be taken into consideration while a combination of official statistics should be used at the macro level. In this way, it is possible to evaluate the multi-dimensional concepts of quality education (SDG 4) together.

This paper has desired to find an answer to the question of whether the government expenditure on education affects the employment rates at ISCED education levels. To tackle this issue, we examine the educational economy efficiency of OECD countries using the relational network DEA, which is a sub-branch of the network DEA model, in order to provide support to policymakers, international education statistics users and academic studies and to determine the indicators that affect quality education. Traditional DEA models perform better than parametric methods in the performance measurement of individual decision-making units. For this reason, it is more accurate to use the traditional DEA based approaches in the research of regional and national education systems and in measuring the performance of the educational economy. However, traditional DEA models are not suitable for measuring the efficiency of substages structures because the performance of interactive substages is neglected. In contrast to traditional DEA models, the relational network DEA can present a systematic view which reflects the countries' correct rank, and provide information about the countries' positioning with regard to indicators used. This analysis shed new light on measuring the educational economy efficiency by taking into consideration indicators on the substages. In this context, we have investigated multistage efficiency scores across the OECD countries by assessing the outputs PISA science performance (stage-1), government expenditure per secondary student (stage-1), employment rate for upper secondary level (stage-2), government expenditure per tertiary student (stage-2), employment rate for tertiary level (stage-3) against inputs directly used in the education system (Government expenditure per primary student (stage-1), PISA science performance (stage-2), Government expenditure per secondary student (stage-2), employment rate for upper secondary level (stage-3), government expenditure per tertiary student (stage-3). By means of having the efficiency of the substages, it was also possible to examine the effects of the indicators used in each substages on the overall educational economy efficiency.

As a consequence of the relational network DEA model's solution, a low-efficiency score is assigned to inadequate units, namely countries, and a high-efficiency score is assigned to adequate units. This efficiency scores reflect the distance to other units in the efficient border estimated during the performance evaluation phase. Thus, the minimum proportional decrease in the inputs or the maximum proportional increase in the outputs of the efficient units can be determined. The empirical results demonstrate that the countries with high-efficiency scores are clustered around countries like Latvia, Slovenia, Korea, and Poland in both overall efficiency and the substages efficiency. In other respect, the countries with low-efficiency scores are clustered around a small number of core countries like Czechia, Mexico, Turkey, and Chile. Therefore, the current paper points out that the relational network DEA can be applied for measuring the educational economy efficiency of the countries due to the capability of providing realistic findings in the country assessment. Besides, it can be said that the relational network DEA models, which provide a scientifically objective analysis and capture the performance complexity of the units dealt with by their nature, are used as an important tool in making international comparisons of country performance in specific areas such as competitiveness, globalization, innovation, and sustainable development. Considering the efficiency scores obtained with this model, the substages efficiencies of the countries define the performance of macroeconomic indicators affecting the education economy at a disaggregated level and enables the analysis of policy areas. On the other hand, the overall efficiency scores of the countries can help determine the policy priorities by determining the extent to which the national performance expectation is met through an international comparison. In this context, network DEA models analyze economic performance beyond simple one-dimensional models that allow analysis between different areas.

ORCID

Deniz Koçak https://orcid.org/0000-0002-5893-0564
Hasan Türe https://orcid.org/0000-0002-1975-9063

5. REFERENCES

Abbott, M., & Doucouliagos, C. (2003). The efficiency of Australian universities: a data envelopment analysis. *Economics of Education Review, 22,* 89 - 97. https://doi.org/10.1016/S0272-7757(01)00068-1

Afonso, A., & Aubyn, M. S. (2005). Non-parametric approaches to education and health efficiency in OECD countries. *Journal of Applied Economics, 8,* 227-246.

Afonso, A., & Aubyn M. S. (2006). Cross-country efficiency of secondary education provision: A semi-parametric analysis with non-discretionary inputs. *Economic Modelling 23,* 476-491.

Afonso, A., Schuknecht, L., & Tanzi, V. (2005). Public sector efficiency: An international comparison. *Public Choice, 123,* 321-347.

Aristovnik, A. (2012). *The impact of ICT on educational performance and its efficiency in selected EU and OECD countries: A non-parametric analysis*. Available at SSRN: https://ssrn.com/abstract=2187482 or http://dx.doi.org/10.2139/ssrn.2187482

Aubyn, M. S. (2003). Evaluating efficiency in the Portuguese education sector. *Economia, 26,* 25-51.

Barra, C., Lagravinese, R., & Zotti, R. (2018). Does econometric methodology matter to rank universities? An analysis of Italian higher education system. *Socio-Economic Planning Sciences, 62,* 104-120.

Chen, Z., & Wu, Y. (2007). The relationship between education and employment: A theoretical analysis and empirical test. *Frontiers of Economics in China, 2* (2), 187-211. https://doi.org/10.1007/s11459-007-0010-4

Ciro, J. A., & Garcia, A. T. (2018). Economic efficiency of public secondary education expenditure: How different are developed and developing countries?. *Desarrollo Sociedad, 80,* 119-154.

Domar, E. D. (1946). Capital expansion, rate of growth, and employment. *Econometrica, 14* (2), 137-147.

Griggs, D., Stafford-Smith, M., Gaffney, O., Rockström, J., Öhman, M. C., Shyamsundar, P., Steffen, W., Glaser, G., Kanie, N., & Noble, I. (2013). Policy: Sustainable development goals for people and planet. *Nature, 495,* 305-307.

Guironnet, J. P., & Peypoch, N. (2018). The geographical efficiency of education and research: The ranking of U.S. universities. *Socio-Economic Planning Sciences, 62,* 44-55.

Gupta, S., & Verhoeven, M., (2001). The efficiency of government expenditure experiences from Africa. *Journal of Policy Modelling, 23,* 433-467. https://doi.org/10.1016/S0161-8938(00)00036-3

Hanushek, E. A., & Kimko, D. D. (2000). Schooling, labor force quality, and the growth of nations. *American economic review, 90* (5), 1184-1208.

Hopkins, C., & McKeown, R. (2002). Education for sustainable development: An international perspective. In Tilbury, D., Stevenson, R. B, Fien J. & Schreuder, D. (Eds.), *Education and Sustainability: Responding to the global challenge,* Commission on Education and Communication, IUCN, Gland, Switzerland and Cambridge, UK.

Jafarov, E., & Gunnarsson, V. (2008). Government spending on health care and education in Croatia: Efficiency and reform options. *International Monetary Fund, IMF Working Paper*, WP/08/136.

Johnes, J. (2006). Data envelopment analysis and its application to the measurement of efficiency in higher education. *Economics of Education Review, 25* (3), 273-288. https://doi.org/10.1016/j.econedurev.2005.02.005

Johnes, J., Portela, M., & Thanassoulis, E. (2017). Efficiency in education. *Journal of the Operational Research Society, 68*, 331 - 338. https://doi.org/10.1016/j.econedurev.2005.02.005

Kao, C. (2009). Efficiency decomposition in network data envelopment analysis: A relational model. *European Journal of Operational Research, 192,* 949-962. https://doi.org/10.1016/j.ejor.2007.10.008

Kashim, R., Kasim, M. M., & Abd Rahman, R. (2017). Measuring efficiency of a university faculty using an extended hierarchical network dea model: A framework. *Advanced Science Letters, 23* (9), 9090-9093

Landau, D. (1983). Government expenditure and economic growth: A cross-country study. *Southern Economic Journal*, *49* (3), 783-792.

Lavrinovicha, I., Lavrinenlo, O., & Teivans-Treinovskis, J. (2015). Influence of education on unemployment rate and incomes of residents. *Procedia Social and Behavioral Sciences, 174* (12), 3824-3831. https://doi.org/10.1016/j.sbspro.2015.01.1120

Le Blanc, D. (2015). Towards integration at last? The sustainable development goals as a network of targets. *Sustainable Development, 23*, 176-187. doi: 10.1002/sd.1582

Lee, J.-W., & Barro, R. J. (2001). Schooling quality in a cross-section of countries. *Economica*, *68* (272), 465-488.

OECD (2012). Portugal, Country Note, Education at a Glance 2012: OECD Indicators Publishing, doi: 10.1787/eag-2012-en

Qin, X., & Du, D. (2018). Measuring universities' R&D performance in China's provinces: A multistage efficiency and effectiveness perspective. *Journal Technology Analysis & Strategic Management*, *30*(12), 1392-1408. doi: 10.1080/09537325.2018.1473849

Ramzi, S., Afonso, A. & Ayadi, M. (2016). Assessment of efficiency in basic and secondary education in Tunisia: A regional analysis. *International Journal of Educational Development*, *51*, 62-76.

Riddell, A. R. (1993). The evidence on public/private educational trade-offs in developing countries. *International Journal of Educational Development*, *13* (4), 373-386.

Ruggiero, J. (2006). Measurement error, education production and data envelopment analysis. *Economics of Education Review*, *25*, 327 - 333. https://doi.org/10.1016/j.econedurev.2005.03.003

Sterling, S. (2001). *Sustainable Education: Re-Visioning Learning and Change. Schumacher Briefings.* Schumacher UK, CREATE Environment Centre, Seaton Road, Bristol, BS1 6XN, England.

Sylwester, K. (2002). Can education expenditures reduce income inequality? *Economics of Education Review*, *21* (1), 43–52. https://doi.org/10.1016/S0272-7757(00)00038-8

UNICEF (2016). The state of the world's children, a fair chance for every child. Available online: https://www.unicef.org/publications/files/UNICEF_SOWC_2016.pdf (9 July 20 18)

United Nations (2015). Transforming our world: The 2030 agenda for sustainable development. Available online: http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E (5 December 2018)

United Nations (2018). Sustainable development goal 4 and its targets, Available online: https://www.un.org/sustainabledevelopment/education (10 November 2018).

Wasylenko, M., & McGuire, T. (1985). Jobs and taxes: The effect of business climate on states' employment growth rates. *National Tax Journal*, *38*(4), 497-511.

Worthington, A. C. (2001). An empirical survey of frontier efficiency measurement techniques in education. *Education Economics, 9* (3), 245-268. doi: 10.1080/09645290110086126

Yang, G.-L., Fukuyama, H., & Song Y.-Y. (2018). Measuring the inefficiency of Chinese research universities based on a two-stage network DEA model. *Journal of Informetrics, 12* (1), 10-30. https://doi.org/10.1016/j.joi.2017.11.002

APPENDIX

Appendix A. *GAMS code to calculate the overall efficiency*

```
SETS
j 'number of DMUS' /DMU1*DMU30/
i 'number of inputs' /X1/
r 'number of outputs' /Y1/
p 'number of intermediates1' /Z1, Z2/
l 'number of intermediates2' /T1, T2/;

TABLE X(j, i) "input matrix"
           X1
DMU1       18.5932
DMU2       23.4497
DMU3       22.4726
DMU4       15.0754
DMU5       15.5312
DMU6       25.6117
DMU7       21.4939
DMU8       21.0185
DMU9       18.0165
DMU10      17.9128
DMU11      18.4491
DMU12      24.4524
DMU13      16.7971
DMU14      21.4471
DMU15      21.3017
DMU16      23.9656
DMU17      31.2500
DMU18      14.8592
DMU19      18.6102
DMU20      19.9855
DMU21      26.4191
DMU22      23.6089
DMU23      19.4372
DMU24      28.8696
DMU25      17.7485
DMU26      25.3405
DMU27      25.3405
DMU28      13.3391
DMU29      22.8298
DMU30      19.8534;

TABLE Y(j, r) "output matrix"
          Y1
DMU1       83.1767
DMU2       85.4733
DMU3       84.4900
DMU4       84.1950
DMU5       84.7400
DMU6       86.2533
DMU7       84.2533
DMU8       83.3400
DMU9       84.0300
DMU10      87.9900
DMU11      81.6667
DMU12      91.2133
DMU13      81.1100
DMU14      85.9100
DMU15      78.1000
```

```
DMU16      77.4733
DMU17      85.1000
DMU18      80.3367
DMU19      86.8167
DMU20      89.5500
DMU21      86.0367
DMU22      82.2067
DMU23      79.9067
DMU24      83.8000
DMU25      77.3800
DMU26      89.1367
DMU27      88.1067
DMU28      76.4267
DMU29      84.9200
DMU30      80.5533;

TABLE Z(j, p) "intermediate1 matrix"
           Z1          Z2
DMU1       509.9939    16.7917
DMU2       495.0375    27.3519
DMU3       501.9997    25.7987
DMU4       446.9561    15.1517
DMU5       492.8300    23.5477
DMU6       501.9369    28.2276
DMU7       534.1937    20.1925
DMU8       530.6612    27.2000
DMU9       494.9776    26.8059
DMU10      509.1406    23.4938
DMU11      476.7475    19.5469
DMU12      473.2301    18.3453
DMU13      502.5751    21.6000
DMU14      466.5528    16.9659
DMU15      480.5468    23.0962
DMU16      515.8099    23.3231
DMU17      490.2250    29.6703
DMU18      415.7099    16.4109
DMU19      513.3035    22.3160
DMU20      498.4811    24.3500
DMU21      501.4353    21.7531
DMU22      501.1001    15.1691
DMU23      460.7749    18.7881
DMU24      512.8636    25.5421
DMU25      492.7861    22.4697
DMU26      493.4224    24.6998
DMU27      505.5058    25.4500
DMU28      425.4895    14.7689
DMU29      509.2215    22.6604
DMU30      496.2424    22.5901;

TABLE T(j, l) "intermediate2 matrix"
           T1           T2
DMU1       77.6700      22.5295
DMU2       75.9433      36.1737
DMU3       72.8533      33.0074
DMU4       71.7000      17.3957
DMU5       77.7067      21.6123
DMU6       79.6433      44.6629
DMU7       75.7733      27.7667
DMU8       73.2000      35.4852
DMU9       72.8867      35.0586
```

```
DMU10     79.4667     37.5885
DMU11     71.5033     23.8811
DMU12     87.0433     25.6424
DMU13     67.6133     25.2509
DMU14     72.4467     19.4927
DMU15     69.8567     26.1989
DMU16     71.9867     13.7440
DMU17     70.9233     22.9545
DMU18     70.6100     40.4662
DMU19     80.9400     27.9934
DMU20     81.2633     38.0238
DMU21     66.1867     24.8147
DMU22     77.4400     25.4516
DMU23     71.2133     20.7731
DMU24     69.5467     21.1539
DMU25     66.0233     22.6820
DMU26     84.1767     43.4855
DMU27     81.2233     38.1637
DMU28     61.8800     24.2958
DMU29     79.5733     37.0920
DMU30     68.0967     24.6532;

parameters
Xo(i) "input vector of DMUo"
Yo(r) "outputput vector of DMUo"
Zo(p) "intermediate1 vector of DMUo"
To(l) "intermediate2 vector of DMUo";

variables
thetaall “efficiency score all”
v(i) "input weights"
u(r) "output weights"
w(p) "intermediate1 weights"
q(l) "intermediate2 weights";

free variables
thetaall;

positive variables
v(i)
u(r)
w(p)
q(l);

equations
EQA
EQB
EQC
EQD
EQE
EQF
EQG
OBJ;

EQA.. SUM (i, v(i) * Xo(i)) =E= 1;
EQB (j).. SUM (r, u(r) * Y(j, r)) – SUM (i, v(i) * X(j, i)) =L= 0;
EQC (j).. SUM (l, q(l) * T(j, l)) – SUM (i, v(i) * X(j, i)) =L= 0;
EQD(j).. SUM (p, w(p) * Z(j, p)) – SUM (i, v(i) * X(j, i)) =L= 0;
EQE (j).. SUM (r, u(r) * Y(j, r)) – SUM (p, w(p) * Z(j, p)) =L= 0;
EQF (j).. SUM (r, u(r) * Y(j, r)) – SUM (l, q(l) * T(j, l)) =L= 0;
```

```
EQG (j).. SUM (l, q(l) * T(j, l)) – SUM (p, w(p) * Z(j, p)) =L= 0;

OBJ.. thetaall =E= SUM (r, u(r) * Yo(r));

*-------------------------------------------
* overall efficiency score
*-------------------------------------------

model overall /
EQA
EQB
EQC
EQD
EQE
EQF
EQG
OBJ
/;

ALIAS (j,o);

LOOP (o,

          LOOP (i, Xo(i) = X(o, i));
          LOOP (r, Yo(r) = Y(o, r));
          LOOP (l, To(l) = T(o, l));
          LOOP (p, Zo(p) = Z(o, p));

          SOLVE overall USING LP maximizing thetaall;
);
```

Appendix B. *GAMS code to calculate the substages efficiencies*

```
SETS
j 'number of DMUS' /DMU1*DMU30/
i 'number of inputs' /X1/
r 'number of outputs' /Y1/
p 'number of intermediates1' /Z1, Z2/
l 'number of intermediates2' /T1, T2/
m 'number of theta3' /thetaall/;

TABLE X(j, i) "input matrix"
            X1
DMU1       18.5932
DMU2       23.4497
DMU3       22.4726
DMU4       15.0754
DMU5       15.5312
DMU6       25.6117
DMU7       21.4939
DMU8       21.0185
DMU9       18.0165
DMU10      17.9128
DMU11      18.4491
DMU12      24.4524
DMU13      16.7971
DMU14      21.4471
DMU15      21.3017
DMU16      23.9656
```

```
DMU17      31.2500
DMU18      14.8592
DMU19      18.6102
DMU20      19.9855
DMU21      26.4191
DMU22      23.6089
DMU23      19.4372
DMU24      28.8696
DMU25      17.7485
DMU26      25.3405
DMU27      25.3405
DMU28      13.3391
DMU29      22.8298
DMU30      19.8534;

TABLE Y(j, r) "output matrix"
            Y1
DMU1       83.1767
DMU2       85.4733
DMU3       84.4900
DMU4       84.1950
DMU5       84.7400
DMU6       86.2533
DMU7       84.2533
DMU8       83.3400
DMU9       84.0300
DMU10      87.9900
DMU11      81.6667
DMU12      91.2133
DMU13      81.1100
DMU14      85.9100
DMU15      78.1000
DMU16      77.4733
DMU17      85.1000
DMU18      80.3367
DMU19      86.8167
DMU20      89.5500
DMU21      86.0367
DMU22      82.2067
DMU23      79.9067
DMU24      83.8000
DMU25      77.3800
DMU26      89.1367
DMU27      88.1067
DMU28      76.4267
DMU29      84.9200
DMU30      80.5533;

TABLE Z(j, p) "intermediate1 matrix"
            Z1          Z2
DMU1       509.9939    16.7917
DMU2       495.0375    27.3519
DMU3       501.9997    25.7987
DMU4       446.9561    15.1517
DMU5       492.8300    23.5477
DMU6       501.9369    28.2276
DMU7       534.1937    20.1925
DMU8       530.6612    27.2000
DMU9       494.9776    26.8059
DMU10      509.1406    23.4938
```

```
DMU11      476.7475      19.5469
DMU12      473.2301      18.3453
DMU13      502.5751      21.6000
DMU14      466.5528      16.9659
DMU15      480.5468      23.0962
DMU16      515.8099      23.3231
DMU17      490.2250      29.6703
DMU18      415.7099      16.4109
DMU19      513.3035      22.3160
DMU20      498.4811      24.3500
DMU21      501.4353      21.7531
DMU22      501.1001      15.1691
DMU23      460.7749      18.7881
DMU24      512.8636      25.5421
DMU25      492.7861      22.4697
DMU26      493.4224      24.6998
DMU27      505.5058      25.4500
DMU28      425.4895      14.7689
DMU29      509.2215      22.6604
DMU30      496.2424      22.5901;

TABLE T(j, l) "intermediate2 matrix"
            T1            T2
DMU1       77.6700       22.5295
DMU2       75.9433       36.1737
DMU3       72.8533       33.0074
DMU4       71.7000       17.3957
DMU5       77.7067       21.6123
DMU6       79.6433       44.6629
DMU7       75.7733       27.7667
DMU8       73.2000       35.4852
DMU9       72.8867       35.0586
DMU10      79.4667       37.5885
DMU11      71.5033       23.8811
DMU12      87.0433       25.6424
DMU13      67.6133       25.2509
DMU14      72.4467       19.4927
DMU15      69.8567       26.1989
DMU16      71.9867       13.7440
DMU17      70.9233       22.9545
DMU18      70.6100       40.4662
DMU19      80.9400       27.9934
DMU20      81.2633       38.0238
DMU21      66.1867       24.8147
DMU22      77.4400       25.4516
DMU23      71.2133       20.7731
DMU24      69.5467       21.1539
DMU25      66.0233       22.6820
DMU26      84.1767       43.4855
DMU27      81.2233       38.1637
DMU28      61.8800       24.2958
DMU29      79.5733       37.0920
DMU30      68.0967       24.6532;

TABLE thetaall(j, m) "efficiency score matrix"
           thetaall
DMU1       0.0538
DMU2       0.0426
DMU3       0.0445
DMU4       0.0663
```

```
DMU5     0.0644
DMU6     0.0390
DMU7     0.0465
DMU8     0.0476
DMU9     0.0555
DMU10    0.0558
DMU11    0.0542
DMU12    0.0409
DMU13    0.0595
DMU14    0.0466
DMU15    0.0469
DMU16    0.0417
DMU17    0.0320
DMU18    0.0673
DMU19    0.0537
DMU20    0.0500
DMU21    0.0379
DMU22    0.0424
DMU23    0.0514
DMU24    0.0346
DMU25    0.0563
DMU26    0.0395
DMU27    0.0395
DMU28    0.0750
DMU29    0.0438
DMU30    0.0504;

parameters
Xo(i) "input vector of DMUo"
Yo(r) "outputput vector of DMUo"
Zo(p) "intermediate1 vector of DMUo"
To(l) "intermediate2 vector of DMUo"
thetaallo(m) "efficiency score vector of DMUj";

variables
theta3 "efficiency score of subprocess 3"
v(i) "input weights"
u(r) "output weights"
w(p) "intermediate1 weights"
q(l) "intermediate2 weights";

free variables
theta3;

positive variables
v(i)
u(r)
w(p)
q(l);

equations
EQA
EQB
EQC
EQD
EQE
EQF
EQG
EQH
OBJ;
```

```
EQA..SUM (p, w(p) * Zo(p)) =E= 1;
EQB (m).. SUM (r, u(r) * Yo(r)) – (thetaallo(m) * SUM (i, v(i) * Xo(i))) =E= 0;
EQC (j).. SUM (r, u(r) * Y(j, r)) – SUM (i, v(i) * X(j, i)) =L= 0;
EQD (j).. SUM (l, q(l) * T(j, l)) – SUM (i, v(i) * X(j, i)) =L= 0;
EQE (j).. SUM (p, w(p) * Z(j, p)) – SUM (i, v(i) * X(j, i)) =L= 0;
EQF (j).. SUM (r, u(r) * Y(j, r)) – SUM (l, q(l) * T(j, l)) =L= 0;
EQG (j).. SUM (r, u(r) * Y(j, r)) – SUM (p, w(p) * Z(j, p)) =L= 0;
EQH (j).. SUM (l, q(l) * T(j, l)) – SUM (p, w(p) * Z(j, p)) =L= 0;

OBJ.. theta3 =E= SUM (r, u(r) * Yo(r));

*-------------------------------------------
* subprocess3 efficiency score
*-------------------------------------------

model subprocess3 /
EQA
EQB
EQC
EQD
EQE
EQF
EQG
EQH
OBJ
/;

ALIAS (j,o);

LOOP (o,

            LOOP (i, Xo(i) = X(o, i));
            LOOP (r, Yo(r) = Y(o, r));
            LOOP (l, To(l) = T(o, l));
            LOOP (p, Zo(p) = Z(o, p));
            LOOP (m, thetaallo(m) = thetaall(o, m));

            SOLVE subprocess2 USING LP maximizing theta3;
);
```

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 436–443
https://dx.doi.org/10.21449/ijate.615640

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

An English Version of the Mathematics Teaching Anxiety Scale

Thomas E. Hunt [1,*], Mehmet Hayri Sari [2]

[1] University of Derby, School of Human Sciences, University of Derby, UK.
[2] Nevşehir Hacı Bektaş Veli University, Faculty of Education, Turkey.

ARTICLE HISTORY

Received: 10 May 2019
Revised: 08 August 2019
Accepted: 21 August 2019

KEYWORDS

Mathematics,
Teaching,
Anxiety,
Scale,
MTAS

Abstract: This study represents the implementation of an English version of the Mathematics Teaching Anxiety Scale (MTAS), originally published in Turkey (Sari, 2014). One hundred and twenty-seven primary school teachers from across the U.K. completed the survey, including 74 qualified teachers and 53 trainees. Following item-reduction and factor analysis, the 19-item MTAS was found to have excellent internal consistency (α = .94) and has a two-factor structure. Factor one, labelled Self-Directed Mathematics Teaching Anxiety, includes 12 items pertaining to a teacher's own teaching practice and perceived ability, whereas factor two, labelled Pupil/Student-Directed Mathematics Teaching Anxiety, includes 7 items pertaining to anxiety concerning pupils/students failing assessments or not reaching curriculum/school targets. Pre-service teachers, compared to in-service teachers, self-reported significantly higher overall maths teaching anxiety. Among in-service teachers, there was a significant negative correlation between length of service and maths teaching anxiety. These findings are important in the context of retention issues in newly qualified teachers and the need to support trainees and newer teachers if they experience anxiety related to teaching maths.

1. INTRODUCTION

Mathematics anxiety is a pervasive issue that appears to exist across a range of populations (Hembree, 1990; OECD, 2013) and can be defined as “feelings of tension and anxiety that interfere with the manipulation of numbers and the solving of mathematical problems in a wide variety of ordinary life and academic situations” (Richardson & Suinn, 1972, p. 551). Empirically measuring anxiety pertaining to numbers began in 1958 with the Numerical Anxiety Scale (Dreger & Aiken, 1957). Since then several self-report scales for measuring maths anxiety have been published (e.g. Richardson & Suinn, 1972; Sandman, 1979; Betz, 1978; Plake & Parker, 1982, Hunt, Clark-Carter & Sheffield, 2011). However, these have been developed for use in a general population without much concern for specific contexts or populations. For example, Baloglu and Kocak (2006) observed that students who majored in elementary (primary) education were amongst the most maths anxious in over seven hundred U.S. university students. This echoes earlier observations that pre-service (student) elementary teachers are especially prone to maths anxiety (Hembree, 1990). As such, it may be necessary to focus on specific populations, e.g. teachers and pre-service teachers. A strong relationship

CONTACT: Thomas E. Hunt ✉ t.hunt@derby.ac.uk University of Derby, School of Human Sciences, University of Derby, UK

ISSN-e: 2148-7456 /© IJATE 2019

has been demonstrated between maths anxiety and confidence in teaching maths among pre-service teachers (Bursal & Paznokas, 2006). In an attempt to reduce maths anxiety in female pre-service teachers, Lake and Kelly (2014) observed little change after completion of an early childhood mathematics course; the authors suggest this is indicative of the students' entrenched beliefs about maths and their ability to do maths. It may also be important to draw a distinction between teachers' and pre-service teachers' maths anxiety and their level of anxiety towards teaching maths. Hadley and Dorward (2011) studied these variables in a large sample (N = 692) of elementary school teachers in the U.S. and found a significant, moderate, positive correlation.

Research on maths anxiety in teachers and pre-service teachers is limited, with only a small amount of work having been conducted in the U.K. One study (Jackson, 2008) investigated 31 British student primary school teachers and found only 19% experienced no negative emotional or physical factors when engaged in maths. Jackson also observed that the students had somewhat negative perceptions of maths and 68% indicated a lack of confidence in teaching maths. Relatedly, Isiksal, Curran, Koc and Askun (2009) found a significant negative relationship between maths anxiety and maths self-concept among trainee teachers in the U.S. and Turkey. In a qualitative investigation, Trujillo and Hadfield (1999) interviewed six highly maths anxious pre-service elementary school teachers in the U.S. and analysis revealed several commonalities amongst the pre-service teachers in relation to their negative emotions pertaining to maths. For example, participants shared negative experiences of maths at school, referring to pressure, poor teaching and humiliation. Similarly, they shared negative experiences of maths within the family, typically referring to unsupportive parents. Shared experiences also extended to magnified anxiety in maths test situations, for example, referring to the maths component of teaching qualification tests. Interestingly, participants expressed a range of attitudes towards teaching maths themselves, seemingly taking into consideration their own negative experiences when planning their lessons; they emphasised previously or currently worrying about preparation and generally advocated a more progressive approach to teaching maths. It may be necessary to consider a range of demographic or individual differences, though. For example, in a further study of pre-service elementary school teachers, Hadfield and McNeil (1994) found a significant positive correlation between age and maths anxiety, such that older participants experienced greater maths anxiety. The authors suggest this may be associated with a lack of confidence in returning students, perhaps due to feeling "rusty" or having a poor background in maths. Providing some support for this argument, Isiksal et al. (2009) found pre-service teachers in the U.S to report significantly higher maths anxiety than pre-service teachers in Turkey, with the authors suggesting the difference might be explained by higher levels of maths familiarity and academic preparedness among Turkish pre-service teachers. Length of time in service may act as a buffer against maths anxiety though, with Gresham (2018) observing a significant reduction in self-reported maths anxiety among ten in-service elementary school teachers five years into teaching.

Interestingly, research findings have indicated it is anxiety towards teaching maths that predicts the adoption of a more traditional teaching style (Hadley & Dorward, 2011). Similarly, Sari and Aksoy (2016) found a negative relationship between maths teaching anxiety and teaching style in Turkey; primary school teachers were found to shift from student-centred teaching to teacher-centred teaching when their mathematics teaching anxiety increased. However, recognised scales that have been developed to specifically test maths teaching anxiety are limited. One scale, the Mathematics Teaching Anxiety Scale (MTAS) (Sari, 2014), was originally developed using a Turkish population of elementary school teachers, although the extent and nature of maths teaching anxiety in the U.K. is unknown. As such, a study using an English version of the MTAS would provide some much needed information, particularly in the context of a poor retention rate of early-career teachers in science, maths and languages (Worth & De Lazzari, 2017) and the need to better understand the reasons for this. Understanding anxiety pertaining to the teaching of maths may also support further research concerning transference of anxiety to students as well as its relationship with teaching style.

2. METHOD

A cross-sectional approach was taken in which an online survey was provided to primary school teachers across the United Kingdom via opportunity sampling.

2.1. Participants

Teachers were required to have (or be working towards) qualified teacher status (QTS). One hundred and twenty-seven participants provided full data, which included 102 (80.30%) females and 25 (19.7%) males. Participant ages ranged from 18 to 69 years (M = 33.57, SD = 12.31) and the sample included 74 (58.30%) qualified teachers (mean age = 40.93 years, SD = 9.96; mean teaching years = 14.97, SD = 9.69) and 53 (41.70%) trainees (mean age = 23.30 years, SD = 6.57).

2.2. Data Collection Techniques

The Mathematics Teaching Anxiety Scale (MTAS) was originally developed using a Turkish population of elementary (primary) school teachers. The MTAS was published in 2014 and contains 23 items. It has high internal consistency (α = .89) and original analyses indicated a three-factor structure (Sari, 2014): i) anxiety regarding maths teaching processes, ii) anxiety regarding maths content knowledge, and iii) anxiety related to maths self-efficacy. The scale lists a range of statements pertaining to different aspects of maths teaching and requires participants to respond on a Likert-type scale regarding how frequently they experience the content of each statement. The response format has five points and ranges from “always” to “never”, with higher scores representing lower anxiety (thus requiring reversing upon data analysis). The scale was originally published in Turkish, so a process of forwards-backwards translation took place, involving multiple academic colleagues, to arrive at an English version of the scale.

2.3. Data Collection Procedure

The survey was administered using Qualtrics online survey software and was advertised via email and social media. Demographic questions were presented first, followed by the maths teaching anxiety measure. Ethical considerations were consistent with the guidelines proposed by the British Psychological Society.

2.4. Data Analysis

The standard procedure was followed, in which internal consistency of the scale and scale items was assessed, followed by an exploratory factor analysis and scale refinement. Group comparisons were made on the teacher variables of sex and teaching status.

3. RESULTS

3.1. Internal consistency – stage 1

The minimum item-total correlation was .42, with a mean of .66. Cronbach’s alpha was .944 and no items were suggested for removal. A Kolmogorov-Smirnov test indicated total scale scores to be significantly positively skewed ($p < .001$); however, inspection of the histogram indicated only slight positive skew.

3.2. Exploratory factor analysis

As the study represents the first administration of an English version of the MTAS, an exploratory factor analysis was conducted. A high Kaiser-Meyer-Olkin measure (KMO = .911) indicated that sampling adequacy was met and low values in the diagonal of the anti-image correlation matrix provided further evidence that the data were suitable for factor analysis (Tabachnick & Fidell, 2001). The mean correlation between extracted factors, based on eigenvalues above one, was <.1, thus indicating independence of factors and therefore verifying the decision to use a varimax rotation. Initially, using eigenvalues above one as criteria for factor extraction, four factors were extracted. The four factors explained a total of 66.56% of the variance, with 46.21%, 10.47%, 5.10%, and 4.74% of the total variance, being explained by factors one to four respectively. The rotated factor matrix revealed several items that did not load sufficiently on to a single factor. In addition, observation of the scree plot indicated the existence of two factors. As such, a further factor analysis was performed in which a two-factor solution was forced. This revealed a much more parsimonious structure in which every item had a factor loading of at least .4. The two factors explained a total of 56.73% of the variance, with 46.26%, 10.47% of the total variance, being explained by factors one and two respectively. Four items were removed due to cross-factor loading. Cronbach's alpha for the resultant 19-item MTAS was .93.

3.3. Factor Labelling

The two factors appeared to represent very distinct underlying constructs pertaining to maths teaching anxiety. In addition to the authors, four independent academics working in the field of maths education were consulted to interpret the nature of factors that the items represent. There was consensus in interpretations. Factor one contained 12 items that relate to a teacher's own teaching practice and perceived maths ability, e.g. "I avoid talking about mathematics teaching with other teachers outside the classroom" and "I worry that I won't be able to answer a question whilst teaching a maths class". Therefore, factor one was labelled Self-Directed Mathematics Teaching Anxiety. Factor two comprised 7 items that relate to teachers' anxiety concerning their pupils, e.g. "The thought that students/pupils will not meet curriculum/school targets in maths worries me" and "I worry that students/pupils in my maths class will fail their assessments". Thus, factor two was labelled Pupil/Student-Directed Mathematics Teaching Anxiety.

3.4. Group comparisons

There was no significant difference between males and females in overall mathematics teaching anxiety, t(125) = 1.27, p = .21, d = 0.28, self-directed mathematics teaching anxiety, t(125) = 1.45, p = .15, d = 0.35, or pupil/student-directed mathematics teaching anxiety, t(125) = 0.42, p = .68, d = 0.09. However, pre-service teachers, compared to in-service teachers, self-reported significantly higher overall maths teaching anxiety, t(125) = 5.78, p < .001, d = 1.07, self-directed mathematics teaching anxiety, t(125) = 6.59, p < 001, d = 1.18, and pupil/student-directed mathematics teaching anxiety, t(125) = 2.12, p = .04, d = 0.38. Among in-service teachers, there was a significant negative correlation between length of service and overall maths teaching anxiety, r(72) = -.27, p = .02, and self-directed mathematics teaching anxiety, r(72) = -.31, p < .01, but not pupil/student-directed mathematics teaching anxiety, r(72) = -.11, p = .38. Means and standard deviations can be seen in Table 1.

Table 1. Means (and standard deviations) of maths teaching anxiety (and sub-scales) according to sex and teacher status.

		Maths anxiety	Self-directed mathematics teaching anxiety	Pupil/student-directed teaching mathematics anxiety
Sex	Males	2.17 (0.50)	1.73 (0.62)	2.94 (0.61)
	Females	2.36 (0.71)	1.99 (0.87)	3.00 (0.67)
Teacher status	Pre-service	2.69 (0.64)	2.44 (0.80)	3.13 (0.55)
	In-service	2.06 (0.58)	1.59 (0.66)	2.88 (0.55)
Overall		2.33 (0.68)	1.94 (0.83)	2.99 (0.65)

4. DISCUSSION and CONCLUSION

This study used an English version of the Mathematics Teaching Anxiety Scale (Sari, 2014) to assess maths teaching anxiety in pre-service and in-service primary school teachers across the U.K. Results suggested a different factor structure to that reported by Sari (2014), including a reduced number of items. Two factors were labelled self-directed mathematics teaching anxiety (12 items) and pupil/student-directed mathematics teaching anxiety (7 items). Respectively, these relate to oneself, including anxiety about one's own maths knowledge, and anxiety directed towards the teaching of others, including worry about one's pupils/students failing assessments or not reaching targets; this second factor is perhaps especially relevant in today's assessment-focused schools.

We observed no significant difference in maths teaching anxiety as a function of sex, which reflects research findings in the field (e.g. Peker & Halat, 2008; Peker & Ertekin, 2011). However, we found pre-service teachers to have significantly higher maths teaching anxiety. Further to this, our results showed that length of service as a qualified teacher was inversely related to maths teaching anxiety. Of note though, this relationship was specific to self-directed mathematics teaching anxiety, suggesting that experience may act as a buffer against anxiety concerning one's own teaching ability regarding maths, possibly due to an increase in confidence. This finding may be particularly important given the previous finding that the higher chance of leaving the teaching profession among younger teachers is the result of inexperience rather than being young. Indeed, primary school teachers with less than 2 years' experience are 5%-10% more likely to leave the profession than those with 6-10 years' experience (Worth, De Lazzari & Hillary, 2017). It is worth highlighting that mean maths teaching anxiety scores in the current study were highest on the pupil/student-directed factor, suggesting particular attention should be paid to teachers' anxiety derived from concerns about pupil/student maths understanding and performance. Indeed, items with the greatest factor loading pertained to anxiety about pupils/students not meeting curriculum/school targets and failing assessments.

A notable consideration is that we studied maths teaching anxiety, not general maths anxiety. As such, the results offer several important points to consider. Firstly, it is reassuring that the overall level of maths teaching anxiety was reasonably low; the mean for the sample represented "rarely" "to sometimes" maths teaching anxious. However, it is noteworthy that the sample includes only those individuals who have not withdrawn from training or teaching, thus suggesting a higher level of resilience than those who have; further investigation is needed on a sample of trainees or teachers that have not been retained in the profession. Secondly, our sample was diverse in terms of locations; it is unknown what training participants had received and the extent to which institutional policies play a part in experiencing maths teaching anxiety. Relatedly, pre-service teachers likely varied in the length of training they had received at the point of completing the survey. Nevertheless, this is the first study to use a validated measure of maths teaching anxiety with a sample of U.K. pre-service and in-service primary school teachers. The scale is easy to administer and may be useful in identifying at-risk teachers and pre-service teachers; more needs to be done to ensure teachers/trainees are supported and not placed under undue stress with regard to teaching maths. Our findings emphasise the multi-dimensional nature of maths teaching anxiety and demonstrates the need to look at the needs of pre-service and in-service teachers separately. The data showed that approximately 14% of respondents scored above "sometimes" in terms of how much maths teaching anxiety they experience. Given the volume of teachers and trainees within primary education this represents a considerable number of individuals in need of additional support and at-risk of leaving the profession due to excess stress in the domain of teaching maths. Whilst care needs to be taken not to over-generalise the findings, this study provides some much needed information concerning the state of maths teaching anxiety within primary education in the U.K., especially in the context of worsening retention rates.

ORCID

Thomas E. Hunt https://orcid.org/0000-0001-5769-1154
Mehmet Hayri Sari https://orcid.org/0000-0002-7159-2635

5. REFERENCES

Baloglu, M., & Kocak, R. (2006). A multivariate investigation of the differences in mathematics anxiety. *Personality and Individual Differences, 40,* 1325-1335.

Betz, N. E. (1978). Prevalence, distribution, and correlates of maths anxiety in college students. *Journal of Counseling Psychology, 25,* 441-448.

Bursal, M., & Paznokas, L. (2006). Mathematics anxiety and preservice elementary preservice teachers' confidence to teach mathematics and science. *School Science and Mathematics, 106,* 173-180.

Department for Education. (2015). School workforce in England: November 2014. Retrieved from https://www.gov.uk/government/statistics/school-workforce-in-england-november-2014

Dreger, R. M., & Aiken, L. R. (1957). The identification of number anxiety in a college population. *Journal of Educational Psychology, 48,* 344-351.

Gresham, G. (2018). Preservice to inservice: Does mathematics anxiety change with teaching experience? *Journal of Teacher Education, 69,* 90-107.

Hadfield, O. D., & McNeil, K. (1994). The relationship between Myers-Briggs personality type and mathematics anxiety among preservice elementary teachers. *Journal of Instructional Psychology, 21,* 375-384.

Hadley, K. M., & Dorward, J. (2011). The relationship among elementary teachers' mathematics anxiety, mathematics instructional practices, and student mathematics achievement. *Journal of Curriculum and Instruction, 5*(1), 27-44.

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal of Research for Mathematics Education, 21,* 33-46.

Hunt, T. E., Clark-Carter-D., & Sheffield, D (2011). The development and part validation of a U.K. scale for mathematics anxiety. *Journal of Psychoeducational Assessment, 29,* 455-466.

Isiksal, M., Curran, J. M., Koc, Y., & Askun, C. S. (2009). Mathematics anxiety and mathematical self-concept: Considerations in preparing elementary school teachers. *Social Behavior and Personality, 37,* 631-644.

Jackson, E. (2008). Mathematics in student teachers. *Practitioner Research in Higher Education, 2,* 36-42.

Lake, V. E., & Kelly, L. (2014). Female preservice teachers and mathematics: Anxiety, beliefs, and stereotypes. *Journal of Early Childhood Teacher Education*, *35*, 262-275.

OECD. (2013). *PISA 2012 results: Ready to learn: Students' engagement, drive and self-beliefs (Volume III).* PISA, OECD Publishing.

Peker, M., & Ertekin, E. (2011). The relationship between mathematics teaching anxiety and mathematics anxiety. *The New Educational Review, 23*, 213-226.

Peker, M., & Halat, E. (2008). *The pre-service elementary school teachers' mathematics teaching anxiety and gender*, The European Conference on Educational Research, 10-12 September, Goteborg, Sweden.

Plake, B. S., & Parker, C. S. (1982). The development and validation of a revised version of the mathematics anxiety rating scale. *Educational and Psychological Measurement, 42*, 551-557.

Richardson, F. C., and Suinn, R. M. (1972). The mathematics anxiety rating scale. *Journal of Counselling Psychology, 19*, 551-554.

Sandman, R. S. (1979). *Mathematics anxiety inventory: User's manual*. Unpublished manuscript, University of Minnesota, Minnesota Research and Evaluation Center, Minneapolis.

Sari, M. H. (2014). Sınıf öğretmenlerine yönelik matematik öğretimi kaygı ölçeği geliştirme [Developing a mathematics teaching anxiety scale for classroom teachers]. *Elementary Education Online*, *13*(4), 1296-1310. Doi: 10.17051/io.2014.11721

Sari, M. H., & Aksoy, N. C. (2016). Sınıf öğretmenlerinin matematik öğretimi kaygısı ile öğretme stilleri tercihleri arasındaki ilişki [The relationship between mathematics teaching anxiety and teaching style of primary school teachers]. *Turkish Studies-International Periodical for the Languages, Literature and History of Turkish or Turkic, 11(3), 1953-1968. Doi: 10.7827/TurkishStudies.9322*

Tabachnick, B. A., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed). Needham Heights, MA: Allyn and Bacon.

Trujillo, K. M., & Hadfield, O. D. (1999). Tracing the roots of mathematics anxiety through in-depth interviews with preservice elementary teachers. *College Student Journal, 33*, 219-232.

Worth, J. & De Lazzari, G. (2017). *Teacher Retention and Turnover Research. Research Update 1: Teacher Retention by Subject.* Slough: NFER.

Worth, J., De Lazzari, G., & Hillary, J. (2017). *Teacher retention and turnover research: interim report.* Slough: NFER.

APPENDICES

The 19-item Mathematics Teaching Anxiety Scale.

1. The thought of not being able to motivate students to learn maths bothers me.
2. The thought that students find maths too abstract concerns me.
3. The thought that students/pupils will not meet curriculum/school targets in maths worries me.
4. The thought that students/pupils will not pay attention to what I am teaching in maths class worries me.
5. I worry that students/pupils in my maths class will fail their assessments.
6. Differences in students'/pupils' prior knowledge worries me when preparing for maths lessons.
7. I worry that students/pupils will answer maths questions incorrectly.
8. At the end of my maths class, I erase the content on the board so that colleagues can't see.
9. I wait for breaks impatiently when I am in maths classes.
10. I am afraid to go beyond the content of maths textbooks.
11. I avoid talking about mathematics teaching with other teachers outside the classroom.
12. I avoid classroom discussion in case students pose difficult maths questions.
13. I get uneasy knowing that the next lesson is mathematics.
14. I feel nervous when a pre-service/trainee teacher observes my maths teaching.
15. I feel uncomfortable when one of my colleagues comes to my classroom during a maths lesson.
16. I worry that I won't be able to answer a question whilst teaching a maths class.
17. Thinking about how to make use of tools/materials that I don't know how to use in the maths classroom makes me feel anxious.
18. The thought of using concrete tools (e.g. geometry boards, pattern blocks, tangrams, fraction bars) in maths classes worries me.
19. I feel uneasy when students/pupils don't understand mathematical concepts and I have to find/think about alternative methods or strategies to teach them.

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 444–460
https://dx.doi.org/10.21449/ijate.550393

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

Safe Learning Environment Perception Scale (SLEPS): A Validity and Reliability Study

Sayed Masood Haidari [1,*], Fazilet Karakuş [1]

[1] Mersin University, Faculty of Education, Department of Curriculum and Instruction, Mersin, Turkey

ARTICLE HISTORY

Received: 07 April 2019
Revised: 24 July 2019
Accepted: 18 August 2019

KEYWORDS

Psychological safety,
Emotional safety,
Classroom environment,
Relationships,
Validity ,
Reliability

Abstract: The purpose of this study was to develop and cross-validate a measurement scale on students' perception of a psychologically safe learning environment in the Turkish context. Primarily, the scale items underwent two rounds of expert review. Then, a series of item elimination or revisions were performed to improve their relevance to the content domain and their comprehensibility for the target group according to the CVI and modified kappa statistics. The results yielded a strong content validity and clarity of the items. Then, the exploratory factor analysis and parallel analysis were performed based on the data from 556 secondary school students (grade 5-8), which suggested a three-factor solution. The KMO was 0.942 > 0.50 with significant Bartlett test values, $x^2(496)$ = 8295.592, p < 0.001 and the explained total variance was 50.622 %. Each item had a factor loading of > 0.58 with > 0.40 common correlations. To validate this structure, confirmatory factor analysis was conducted based on the data from a different group of students (N = 339). The goodness of fit indices, factor loadings, and the t statistics supported a good-fitting measurement model, $x^2(N = 339) = 925.29$, $df = 461$, $p < 0.001$; $x^2/df = 2$, NFI = 0.94, NNFI = 0.97, CFI = 0.97, SRMR = 0.069, RMSEA = 0.055. The convergent and discriminant validity were also supported. In general, the SLEPS has potential applicability both at the lower and upper secondary schools (public and private) and at the educational centers for the gifted.

1. INTRODUCTION

Nurturing a sense of emotional and psychological safety is essential in every learning environment to facilitate effective teaching and learning opportunities (Holley & Steiner, 2005). Establishment of healthy relationships and positive social interactions in the classroom can be the main prerequisites to start forging an atmosphere of such kind. One thing is for sure that the adolescents today are sensitive to the negative and extreme behaviors, which can easily result in distraction, sense of fear and unlearning thereafter. They "need more emotional and social guidance to cope with social pressure and personal identity" (Beamon, 2001, p. 3). Thus, the classroom must better be ready to ease such pressures as the right place, where individuals are meant to be educated as worthy members of society. In other words, they need guidance and

CONTACT: Sayed Masood Haidari ✉ haidarimasood@gmail.com 🖃 Mersin University, Faculty of Education, Department of Curriculum and Instruction, Çiftlikköy Campus, 33110 Mersin, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

support in promoting their self-confidence, self-esteem and emotional security to be raised as healthy members of the society from a social and psychological perspective. This will be, of course, possible when the classroom population is protected "from psychological or emotional harms" (Holley & Steiner, 2005, p.50). Creating a safe space in the classroom, where student identity and individuality are valued and nourished is essential in enforcing student connectedness to the learning environment. This is what Foldy, Rivard, and Buckley (2009) refer to as *identity safety*. Identity safety means upholding a perception that nobody will despise my social position within a group of learners. Though Foldy et al. (2009) relate the identity safety to a learning environment where individuals come from different racial backgrounds; its vitality is sensed in every academic context that accommodates learners with different backgrounds in terms of their individual needs, unique abilities, personal characteristics and so forth.

The word *safe space* here is a metaphoric attribution to a learning environment, where problematic issues, hard feelings, behavioral problems, and unnecessary pressures are impeded (Gayle, Cortez & Preiss, 2013; Holley & Steiner, 2005). Instead, a sense of connectedness to the classroom and willingness to engage in the activities are nourished in students, getting them convinced that the classroom is a psychologically safe place for learning, "where risks can be taken, mistakes can be made, and understanding can be gained" (Gayle et al., 2013, p. 2). Herewith, students should feel free to share their honest opinions, ask questions and learn enthusiastically without being subjected to embarrassment or humiliation (Turner & Braine, 2015). If they do not feel mentally safe and comfortable, critical thinking will not flourish and the ideas shared will not be real, but fabricated. Raghallaigh and Cunniffe (2013) argue that experiencing uncertainties and sense of fear hinder student involvement in classroom activities. Feeling psychologically unsafe can be among the main reasons that increase uncertainties in students. As a result, they might be concerned about the consequences of giving wrong answers or bringing up questions with a fear that might reflect their ignorance. Beamon (1993) contemplates that establishing a safe learning space in the classroom to improve student thinking ability is interrelated with how "teacher interacts with, responds to and challenges" them by asking well-formulated cognitive questions and facilitating their participation in discussions (p. 91).

Given that self-disclosure, risk-taking, critical thinking, and positive relationships are fostered if individuals feel safe amongst a group of learners within the classroom, then setting up a psychologically safe learning atmosphere is mandatory. To build a safe and welcoming learning environment, everyone in the classroom is supposed to feel secure (Foldy et al., 2009). Besides, students should not get punished or ridiculed for their ways of thinking. Instead, they must be encouraged "to take risks, honestly express their opinions… share and explore their knowledge, attitudes, and behaviors" and make the classroom a safe haven for them to progress in individual level (Holley & Steiner, 2005, p. 50). However, bullying, harassment, and ridicule may seriously hurt students' feelings and negatively affect their learning process, curbing their participation in classroom activities. A safe and desirable learning environment requires creating participative and rich learning opportunities to the students so that they feel connected to their teacher and classmates. Vice versa, having a sense of belonging or connectedness to the classroom improves student participation and participation can promote a feeling of safety and acceptance (Frisby, Berger, Burchett, Herovic & Strawser, 2014). This will, in turn, improve student-student and student-teacher relationships in addition to building a trusted plus respectful learning atmosphere in the classroom. Establishing positive and trusted relationships in the classroom can give the type of morale they need and encourage them to reveal their thoughts instead of protecting their image and individual self from potential embarrassments.

A study with gifted-students and students having emotional or behavioral disorders in an Australian secondary school revealed that teacher behaviors contribute to building positive relationships between the teacher and students, resulting in productive learning opportunities thereafter (Capern & Hammond, 2014). The study further reported that the gifted students valued the friendliness and cordiality of the teachers that ultimately leads to productive learning. However, the students with emotional or behavioral disorders considered the teachers' behavioral characteristics "*that displayed warmth, understanding and patience*" as preconditions to effective learning (Capern & Hammond, p. 46). Treating students with respect, giving equal opportunities for self-disclosure, allowing peer assistance, and not discriminating between them were among the other findings that indicate the importance of approachability of the teacher displayed through his/her behaviors, influencing student learning.

Truly, active student engagement in learning results in improved learning, better academic performance, and personal-growth (Raghallaigh & Cunniffe, 2013; Frisby et al., 2014). Participation, of course, could be reinforced by cultivating a sense of confidence in students and fostering challenging, yet enjoyable learning experiences (Gayle et al., 2013). Thus, encouraging in-class student interactions and building sincere relationships falls to the teachers to control the situation in favor of the students. In the meantime, the approachability of the teacher, his willingness to listen to the student voices, cherishing diverse opinions and attending student needs in individual level are vital in creating a desirable learning space that is psychologically safe (Gillen, Wright, & Spink, 2011).

Nevertheless, safe learning spaces must not be confused with unchallenging and conflict-free environments (Boostrom, 1998; Holley & Steiner, 2005). Where there is no conflict, there is no learning and critical thinking. Here, conflict refers to the diversity of thoughts, conflict of ideas and disagreements as natural parts of learning. The presence of psychological safety "may decrease barriers to engagement and allow individuals…to interact with the world around them" (Wanless, 2016, p. 6). This may persuade students to come out of their comfort zones in order to reveal their individuality by expressing themselves openly and honestly as well as develop their knowledge and skills. If students do not get exposed to academic challenges, they may not progress as required. Boostrom (1998) maintains that "…teachers need to manage conflict, not prohibit it" to flourish "critical thinking" in students (p. 407). To tackle academic challenges and conflictive ideas, mental safety needs to be insured and students should get encouraged to voice their opinions bravely in an academic context (Boostrom, 1998). However, accepting every standpoint without constructive criticisms might hinder personal-growth. Simply put, students must feel emotionally safe in order to be open to critical evaluation of their opinions by other students in the class.

Surely, not being mocked and disgraced because of uncommon ideas, incorrect answers or asking questions differ from alerting students of their ignorance and learning deficiencies (Holley & Steiner, 2005). According to Wanless (2016), discomforts are inevitable when new opinions are shared, but they do not have to get in the way of the students to achieve their goals. Hereby, two main responsibilities fall to the teacher while trying to create a psychologically safe atmosphere: i.e. a) inhibiting annoying acts and bad behaviors in the class that might prevent taking creative risks, and b) informing students of their academic progress without being judgmental and discriminative. If students are judged for what and who they are, it may undermine their learning and push them towards alienation.

However, how the psychologically safe learning climate is perceived might differ from student to student, (i.e. gifted-students from normal students and students with different socio-economic backgrounds) particularly in lower secondary education level. As noted before, middle school students or so-called *adolescents* (Beamon, 1993) are more susceptible to the negative features of psychologically unsafe learning environments. Beamon (1993) conceptualizes that at the secondary "grade level, where students' intellectual capacity is rapidly unfolding", promoting their "thinking ability is a critical one" (p. 92). As it appears, few empirical studies exist (mostly qualitative ones) as regards psychologically safe learning environments. The sample in the existing studies is mostly from universities or colleges. Besides, a valid and reliable quantitative data collection tool was not found about the student perception in this regard. Therefore, this study was conducted to develop a safe learning environment perception scale (SLEPS) suitable to the secondary level students (grade 5-8) from diverse socioeconomic backgrounds studying at public, private and gifted schools.

2. METHOD

2.1. Participants

The data were collected from a total of 651 secondary level students (grade 5-8) for the exploratory factor analysis (EFA). However, this number dropped to 556 after the removal of eight incomplete and 87 multivariate outliers. The sample was selected from different socioeconomic backgrounds studying in the private or public schools and educational centers for the gifted students (i.e. extra schooling that the gifted-students receive in *Science and Art Centers* besides attending public schools) in Turkey (see Table 1). Their mean age was M = 11.67 (SD = 1.28) of both gender. Secondly, another set of data was collected from 349 students for the confirmatory factor analysis (CFA), but this number decreased to 339 after ten univariate or multivariate outliers were removed. The mean age of the students in this group was M = 11.86 (SD = 1.31).

Table 1. Demographic information about the sample in EFA (N = 556) and CFA (N = 339)

Variables	Category	EFA		CFA	
		n	%	*n*	%
School	Public	301	54.1	142	41.9
	Private	203	36.5	102	30.1
	Gifted	52	9.4	95	28
Gender	Male	292	52.5	162	47.8
	Female	264	47.5	177	52.2
Grade Level	Five	191	34.4	121	35.7
	Six	153	27.5	66	19.5
	Seven	100	18	82	24.2
	Eight	112	20.1	70	20.6
	Total	556	100	339	100

2.1. Item Development, Expert Review and Content Validity Index

The items were developed after reviewing relevant literature on safe learning environment. Initially, 90 items were generated. However, this number dropped to 85 after removing five items because of their similarity to the other items. All of them were written in the Turkish language because of the sample characteristics. To ensure the content validity of the items, an expert review form was devised with two criteria of *relevancy* and *clarity* being measured by a four-point scoring system. This form required the experts in the education field to rate the relevance of the items to the content domain and the level of their clarity or comprehensibility as "1=*not relevant*, 2=*somewhat relevant*, 3=*quite relevant*, 4=*highly relevant*" (Pilot, Beck & Owen, 2007, p.460). The same scoring system was applied to the clarity criteria from 1-4, 1 being *not clear* and 4 being *highly clear*. Afterward, two different expert reviews were conducted. At stage one, seven university Ph.D. lecturers with different expertise in the field of education were selected. Two of them were assessment and evaluation experts to evaluate the psychometric properties of the items. After they returned the review forms, the scores were entered in two different Excel tables to calculate the content validity index (CVI) and modified kappa.

CVI shows the “Degree to which an instrument has an appropriate sample of items for construct being measured” (Pilot & Beck, 2006, p. 493) while “modified kappa statistics adjust for chance agreement” to the items agreed to be relevant by the experts; not their agreement on irrelevant ones (Pilot et al., 2007, p. 465). Moreover, items having an item level CVI (I-CVI) or kappa values of 0.80 or more were kept while others below that threshold were removed, though according to Pilot et al. (2007) an I-CVI of 0.78 is acceptable showing adequate content validity with three or more reviewers. However, according to Lynn (1986), the I-CVI should be ‘1’ with three to five reviewers while it can be relaxed when they are more than five. This means that all the reviewers must agree on the relevance or the clarity of the items if they are five or under that number. With the removal of items under I-CVI of 0.80, the number of items dropped to 68. After the recommended revisions, the second round of expert review was conducted with three reviewers, following the same procedure. Since the number of reviewers was three, then the I-CVI of ‘1’ was considered acceptable (Lynn, 1986).

I-CVI was calculated by counting the number of reviewers who rated each item as 3 or 4 to the total number of reviewers. Then the scale level CVI (S-CVI) was calculated in two different ways to check the overall relevancy/clarity levels: a) calculating I-CVI average (S-CVI/AV) and, b) universal agreement of S-CVI (S-CVI/UA). S-CVI/AV was calculated by dividing the total of I-CVI to the total number of the items. However, the S-CVI/UA was computed by dividing the sum of items that received 3 or 4 from all reviewers to the sum of all items. S-CVI/AV and S-CVI/UA of 0.80 or more were considered acceptable (Pilot & Beck, 2006; Pilot et al., 2007). Further, the probability of chance agreement as the prerequisite of the modified kappa (also called k^*) was computed using this formula: “$P_c = [N!/A!(N-A)!]*.5^N$” (Pc = probability of chance agreement, N = number of experts, A = number of experts giving a score of 3 or 4 to an item). After that, k^* was calculated by employing the I-CVI proportion of agreement and Pc through “K = (I-CVI - Pc) / (1 - Pc)” formula (Zamanzadeh et al., 2015, p. 69; Pilot et al., 2007, p. 466).

Followed by two rounds of expert review, major item revisions, application of different inter-rater tests, and elimination of irrelevant items, the last version of the SLEPS was devised containing a total of 59 items in a five-point Likert scale format to be responded accordingly by the selected sample (5= Strongly disagree, 4=Agree, 3=Undecided, 2, Disagree, 1=Strongly Disagree). Demographic information about the participants of the study was sought as regards their school type, grade levels, gender, and age.

2.2. Analysis

After the data were collected, they were screened to identify the incomplete cases. As such eight cases were found and discarded. The remaining data were entered into the SPSS program and the negative items were reverse coded. Then the dataset was screened for univariate and multivariate outliers. For detecting the univariate outliers, the standardized *z* scores were evaluated and for multivariate outliers, the Mahalanobis Distant values were compared to Chi-square Table of the critical values as presented in Tabachnick and Fidell (2013). To ensure the factorability of the data set, the Kaiser-Meyer-Olkin (KMO) and Bartlett Test of Sphericity were run. The KMO value was expected to be over 0.6 and the Bartlett test results to be significant at $p < .001$ level (Aldrich & Cunningham, 2016). The linearity check between the variables, and the “Multicollinearity and singularity” analysis were also conducted as the prerequisites to the exploratory factor analysis (EFA) outlined in Tabachnick and Fidell (p. 674).

After conducting the required tests above, a principal components analysis (PCA) was run followed by a principal axis factoring (PAF) to extract the latent variables for the SLEPS. These analyses were employed "to describe and summarize data by grouping together variables that are correlated" under the extracted latent variables (Tabachnick & Fidell, 2013, p. 614). The size of loading for each variable was decided to be at least 0.45 to be retained. Besides, the eigenvalue greater than '1' was considered acceptable for factor determination. To ensure that the factor determination through eigenvalues and scree plot is not by chance, the Horn's parallel analysis (PA) was run as an alternative objective method in factor determination (Patil, Singh, Mishra, & Donavan, 2008). Sometimes, relying on eigenvalues greater than '1' rule and scree plots can be misleading and lead to over-estimation of factors due to its subjectivity. According to Patil et al. (2008), PA provides information that is more accurate in this regard. It is conducted by comparing the actual eigenvalues of the extracted factors "with eigenvalues extracted from a randomly generated correlation matrix having the same sample size and number of variables", where the eigenvalue in the actual data is expected to be larger than the values estimated in the simulated data for a factor to be retained (Patil et al., 2008, p. 164; Williams et al., 2010; Çokluk & Koçak, 2016). To calculate PA, a "Web-based PA engine", developed by Patil et al. (2007), was used where only the total number of items and the sample size is required to generate the random eigenvalues (Patil et al., 2008, p, 168; see Patil et al., 2007). In addition, to decide on rotation type, the correlations between the factors were evaluated. According to Aldrich and Cunningham (2016), if factors are correlated an oblique rotation technique is used. Otherwise, an orthogonal rotation is preferred. Analyses were repeated after excluding two types of items: a) complex items cross loading on more than one factor with less than 0.10 difference and b) the items below the specified cutoff value for the factor loading (Seçer, 2013).

Furthermore, a CFA was performed, through Maximum Likelihood estimation method in LISREL 8.71, to validate the measurement model for the SLEPS. A series of assumption tests were conducted as done in EFA. The missing values lower than 5 % were imputed via mean substitution. Then the chi-square value (x^2) and the goodness of fit statistics were analyzed and reported by evaluating the resultant values of Root Mean Square Error of Approximation (RMSEA), Normed Fit Index (NFI), Non-Normed Fit Index (NNFI), Comparative Fit Index (CFI), and Standardized Root Mean Square Residual (SRMR). The RMSEA and SRMR values of 0.08 or lower plus the NFI, NNFI, and CFI values of 0.90 or over were considered as adequate model fit indexes (Stevens, 2009; Seçer, 2013, p. 152; Pituch & Stevens, 2016, p. 654). Besides, the normed chi-square was calculated by dividing the x^2 to its degree of freedom, considering the recommended ratio of 3 or smaller (Walts, Strickland, & Lenz, 2010). Next, the Chronbach's alpha, the average variance extracted (AVE), the composite reliability (CR), and the squared correlation between the factors were calculated to provide results for the internal consistency, convergent, and discriminant validity respectively.

3. FINDINGS

3.1. Content Validity Computation Results

The first round of expert review on item relevancy and clarity resulted in major revisions in accordance with the recommendations made. Besides, 17 out of 85 items were eliminated because of receiving low relevancy scores (I-CVI < 0.80, $k^* < 0.80$). The S-CVI/AV test results indicated a high level of overall relevancy of the items (0.916) to the content domain, while S-CVI/UA (0.682) indicated the opposite. In addition, the clarity test results showed that most of the items needed revision, for they were not comprehensible enough. As the item level analyses indicated (I-CVI < 0.80, $k^* < 0.80$), 34 out of 85 items were rated either '1' or '2' needing major revisions to make them conceptually comprehensible to the lower secondary level students. Although S-CVI/AV was found to be at an adequate level (0.818), the S-CVI/UA score was unacceptable (i.e. 0.60). Therefore, after fundamental changes were made to increase the clarity of the items, a second expert review form was devised with 68 items. This time three experts rated the relevancy and clarity of the items once more. Only one item was eliminated because of a low I-CVI (0.66).

However, the results of S-CVI/AV and S-CVI/UA show that the overall relevancy level of the scale items is excellent (0.995 and 0.985 respectively). Similarly, the clarity scores of the items considerably increased considering the S-CVI/AV and S-CVI/UA (0.961 and 0.882 respectively) above the acceptable level although one of the experts rated eight items as ‘2’ meaning not clear enough and thereby suggested some corrections. She also suggested eliminating some of the items because of their similarity to other items. Therefore, the final form of SLEPS was devised with the inclusion of 59 and exclusion of nine items, as they were too similar to some other items in the form in terms of meaning.

3.2. Exploratory Factor Analysis

To extract the potential latent variables for the SLEPS, a PCA followed by a PAF was employed to see how the results compare. However, as a conservative approach in factor extraction, PCA results were prioritized. Prior to these analyses, the required assumption tests were run. As mentioned elsewhere, the dataset was scrutinized for possible univariate and multivariate outliers. The analysis of the standardized *z* scores indicated that no univariate outliers exist. However, the comparison between the Mahalanobis distance values of multiple regression and the Chi-square table of critical values considering the degree of freedom of 59 at $p < .001$ level indicated that 87 cases have a Chi-square of $x^2 = 90.607$ or more, fall into multivariate outliers category and therefore were excluded from the study (Tabachnick & Fidell, 2013, p. 952). This way, the sample size dropped to 556 for the subsequent analysis. Sample sizes of 100, 300, and 1000 are classified as poor, good and excellent respectively (Field, 2009). To ensure the factorability of the data, the KMO sampling adequacy and Bartlett Test of Sphericity were run. The KMO, 0 .942 > 0.50, and the Bartlett test results, $x^2(496) = 8295.592$, $p < 0.001$, supported the suitability of the data for factorization (Williams, Onsman, & Brown, 2010).

Moreover, the data indicated normal distribution according to standardized z scores and the similarity of the central tendency measures (mean, median and mode) across all variables. Besides, because of the impracticability of analyzing pair-wise linearity scatter plots of all the variables in the study, it was decided to check the linearity between the variables holding the most negative and the most positive skewness values. The result showed a nonlinear relationship between the two. Hence, the data were screened for multicollinearity, singularity and auto-correlation problems. According to the evaluation of the coefficients table, the variance inflation factors (VIF) were < 5, the Tolerance values > 0.20 in all cases, and the Durbin-Watson test value was 2. These results prove that the respective problems mentioned do not exist in the data.

Finally, the initial PCA was run after the essential analysis made above. The factor loading for each item was decided to be at least 0.45 to be retained and the eigenvalues were set on ‘1’ as a default acceptable level for factor determination (Tabachnick & Fidell, 2013). These values ranged from 1.001-16.668 suggesting a nine-factor solution with the largest total variances explained by 55.65 %. However, the sub-factor with 5 % explained variances were preferred. When the initial total variances of the eigenvalues were evaluated, only the first three factors met this criterion explaining 42.35 % of total variances. Contrarily, the initial PAF analysis of EFA yielded a relatively different result in terms of the explained total variances both for the nine-factor (47.41 %) and the three-factor solutions (39.71 %). In general, the eigenvalues with the explained variances of 5 % or more, as well as the inflexion point on the scree plot, suggested a three-factor solution (see Figure 1). The factors were not determined only with these two measures but also by considering the PA results (see Table 2).

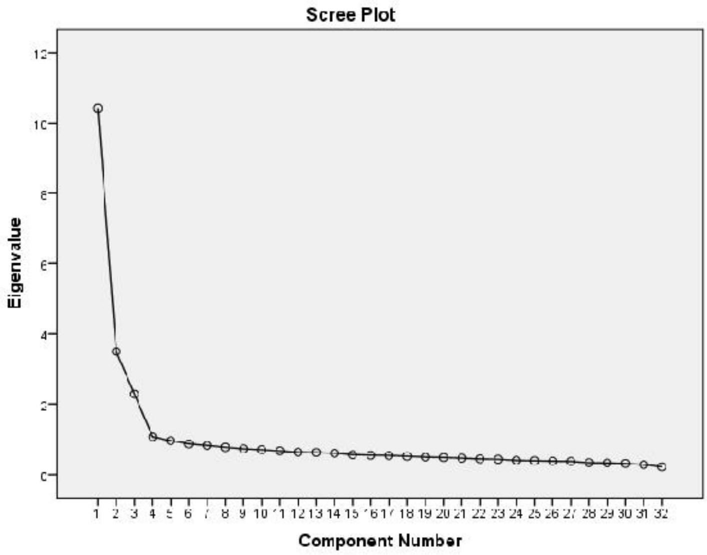


Figure 1. The Scree plot, showing the number of retainable factors

As shown in Table 2, the comparison between the real eigenvalues generated by SPSS and the randomly generated eigenvalues through a web-based parallel analysis (Patil et al., 2007), indicate that the retained factors through PCA is not by chance, because the first three factors hold greater eigenvalues in the real dataset than the random one. However, in the fourth factor the random eigenvalue (1.591141) was greater than that of the real one (1.064), confirming the three-factor solution suggested by SPSS.

Table 2. Eigen values and the explained variances after PCA and PAF

	Eigenvalues		PCA		PAF	
Factors	Real	Random	% of Variance	Cumulative %	% of Variance	Cumulative %
1	10.420	1.753312	32.561	32.561	30.908	30.908
2	3.496	1.686328	10.925	43.487	9.239	40.147
3	2.283	1.637604	7.135	50.622	5.399	45.546

The eigenvalues for the three factors ranged from 2.283 to 10.420 and the explained variances of the individual factors ranged from 7.135 to 32.561 % with explained total variances of 50.622 %. However, in comparison to PCA results, the PAF analysis yielded smaller explained variances, ranging from 5.399 to 30.905, with a cumulative percent of explained total variances of 45.546 % showing a difference of around 5 %.

Considering the eigenvalues and the scree plot, the number of factors was set on three. To decide on rotation type, the correlation between the factors was evaluated. After ensuring they are inter-correlated, an oblique rotation of Promax was preferred. In this regard, Brown (2015) argues, “oblique rotation provides a more realistic representation of” inter-correlation between the factors. However, if they are not correlated, “oblique rotation will produce a solution that is virtually the same as one produced by orthogonal rotation” (Brown, 2015, p. 28). Afterward, the communalities table was evaluated and the items being correlated under .40 were noted for later exclusion. When the rotated component matrix was evaluated several items were found either cross-loading on more than one factor or under cutoff value (0.45) for factor loadings. With the exclusion of these under-correlated and complex items (27 out of 59), a 32-item scale was devised through PCA (See Table 3).

Table 3. The analyses result for the 32-item scale after the PCA with Promax rotation

Items	Factor Loadings				
	F1	F2	F3	h^2	Item-Total Correlations
Item36	.815			.599	.605**
Item45	.783			.599	.655**
Item53	.781			.579	.647**
Item6	.771			.557	.515**
Item30	.758			.530	.560**
Item29	.755			.556	.572**
Item24	.731			.486	.615**
Item34	.693			.523	.629**
Item40	.664			.525	.636**
Item58	.658			.551	.594**
Item48	.631			.505	.675**
Item28	.624			.428	.585**
Item59	.623			.466	.543**
Item18	.615			.404	.562**
Item47	.596			.405	.627**
Item21	.561			.428	.616**
Item19		.799		.558	.542**
Item51		.709		.595	.557**
Item42		.708		.500	.631**
Item8		.695		.475	.441**
Item3		.665		.490	.558**
Item33		.641		.410	.536**
Item41		.639		.449	.593**
Item31		.578		.458	.588**
Item22		.571		.443	.593**
Item55_R			.775	.655	.505**
Item39_R			.728	.507	.347**
Item50_R			.723	.505	.349**
Item44_R			.710	.512	.320**
Item57_R			.677	.534	.447**
Item32_R			.675	.510	.424**
Item46_R			.657	.455	.427**
Explained Variances (50.622 %)	32.561 %	10.925 %	7.135 %		
Cronbach's Alpha	.93	.86	.84		

Note: F1 = Teacher Approachability, F2 = Positive Peer Relationships, F3 = Lack of Identity Safety, $^{**}p < 0.01$

As reported elsewhere, the resultant item elimination process contributed to an increase in the explained total variance from 46.756 % to 50.622 %. Similarly, the explained total variance obtained through PAF also increased from 39.71 % to 45.546 %, which is relatively smaller than the one obtained through PCA (50.622 %).

As illustrated in Table 3, the first factor contains 16 items, the second 9 and the last 7. Besides, the third factor contains only negative items that were reverse-coded and indicated with R letter at the end. The communalities column (h^2) shows that the common correlation value for each item was above 0.40 and the item-total correlations were significant across all variables ($p <$ 0.01), ranging from minimum 0.320 to maximum 0.675. Similarly, the factor loadings were high enough, ranging from 0.561 to 0.815 for the first factor, 0.571 to 0.799 for the second, and 0.657 to 0.775 for the third factor. All the variances explained by each factor are given beneath Table 3 including the Cronbach's Alpha values of reliability. The Cronbach's alpha was .93 for the first factor, 0.86 for the second, and .84 for the third. These results indicate that this measurement scale (i.e. SLEPS) comprising of three factors and 32 items is reliable because of its high internal consistency across all observed variables and their respective latent variables. Then these latent variables were named according to the sub-dimensions of the safe learning environment in the literature. So to speak, the first factor was named as *Teacher Approachability* whereas the second and third were named as *Positive Peer Relationships* and *Lack of Identity Safety* respectively.

3.2. Confirmatory Factor Analysis

A correlated traits model of CFA, based on data from 339 secondary school Turkish students, was performed to cross-validate the three-factor SLEPS developed in the present study. Initially, the negative items were reverse-coded followed by a series of assumption tests. So doing, one univariate and nine multivariate outliers were identified and therefore excluded. Fifteen cases with less than 5 % missing values were imputed through mean substitution. However, there was no issue of concern regarding the singularity and multicollinearity since the VIF (< 5) and Tolerance (> 0.20) values were under the threshold. The model fit indices, estimated through Maximum Likelihood approach, were compared to and interpreted according to the recommended cutoff values, within acceptable ranges, in the literature (Hu & Bentler, 1999, p. 27; Stevens, 2009; Kline, 2016; Pituch & Stevens, 2016, p. 654).

Table 4. The model fit measures for the SLEPS

Fit Indices	Perfect Fit	Adequate Fit	Fit Indices of SLEPS	Model Fit Level
x^2/df	0 or < 2	≤ 3	2	Adequate
NFI	0.95 or close to 1	≥ 0.90	0.94	Adequate
NNFI (TLI)	0.95 or close to 1	≥ 0.90	0.97	Perfect
CFI	0.95 or close to 1	≥ 0.95	0.97	Perfect
SRMR	0 or ≤ 0.050	≤ 0.08	0.069	Adequate
RMSEA	0 or ≤ 0.050	≤ 0.08	0.055	Adequate

From Table 4 it can be seen that fit indices for the measurement model under three factors support a good fit without any modification, x^2 (N = 339) = 925.29, df = 461, p < 0.001; x^2/df = 2, NFI = 0.94, NNFI (TLI) = 0.97, CFI = 0.97, SRMR = 0.069, RMSEA = 0.055. Here, the normed x^2/df, NFI, SRMR, and RMSEA were adequate. However, the NNFI (TLI) and the CFI were at the perfect level since these indexes were very close to 1 (Hu & Bentler, 1999). Similarly, the second order CFA under one factor yielded the same results in terms of fit indices and parameter estimates. Therefore, no modification was performed since the suggested model was of a good fit. Moreover, Table 5 indicates that the standardized parameter estimates (i.e. factor loadings) range from 0.59 to 0.77 for Teacher Approachability, 0.50 to .71 for Positive Peer Relationships and 0.45 to 0.75 for Lack of Identity Safety, all with significant t statistics (p < 0.01). The standardized and unstandardized factor loadings for the individual items, their error terms, and t values can be evaluated in the respective table.

Table 5. Standardized and unstandardized parameter estimates, standard errors, and *t* values

Factors	Items	Unstandardized Estimates	Standardized Estimates	*SE*	*t*
Teacher Approachability	Item36	0.60	0.67	0.55	13.56**
	Item45	0.73	0.64	0.59	12.85**
	Item53	0.79	0.67	0.55	13.63**
	Item6	0.64	0.70	0.50	14.50**
	Item30	0.58	0.61	0.63	12.07**
	Item29	0.48	0.63	0.61	12.47**
	Item24	0.82	0.77	0.41	16.40**
	Item34	0.71	0.68	0.53	13.92**
	Item40	0.68	0.67	0.55	13.65**
	Item58	0.75	0.63	0.61	12.46**
	Item48	0.75	0.63	0.60	12.51**
	Item28	0.71	0.59	0.66	11.49**
	Item59	0.89	0.68	0.53	13.95**
	Item18	0.73	0.65	0.58	13.06**
	Item47	0.69	0.61	0.63	11.95**
	Item21	0.81	0.60	0.63	11.91**
Positive Peer Relationships	Item19	1.03	0.71	0.50	14.12**
	Item51	0.70	0.58	0.66	10.95**
	Item42	0.91	0.70	0.51	13.96**
	Item8	0.59	0.61	0.63	11.62**
	Item3	0.94	0.69	0.53	13.63**
	Item33	0.75	0.50	0.75	9.21**
	Item41	0.80	0.61	0.62	11.73**
	Item31	0.77	0.61	0.63	11.72**
	Item22	0.71	0.50	0.75	9.19**
Lack of Identity Safety	Item55_R	1.13	0.75	0.43	14.82**
	Item39_R	0.95	0.62	0.62	11.53**
	Item50_R	0.77	0.59	0.65	10.88**
	Item44_R	0.82	0.55	0.70	9.92**
	Item57_R	0.51	0.45	0.80	7.87**
	Item32_R	0.97	0.61	0.63	11.27**
	Item46_R	0.82	0.56	0.69	10.10**

Note: ***p* < 0.01

The reliability statistics of Cronbach's Alpha was calculated once again according to the final measurement model. As seen in Table 6, the alpha value was 0.919 for Teacher Approachability, 0.839 for Positive Peer Relationships, and 0.789 for Lack of Identity Safety. These results were similar to the reliability statistics calculated after EFA, showing an adequate level of internal consistency between the latent and observed variables.

In addition, the AVE values of the factors ranged from 0.36 to 0.43, which are under the threshold of 0.50 (see Table 6). However, the CR values were found adequate (above 0.70), ranging from 0.791 to 0.922. The latter findings indicate good reliability and therefore can be accepted as a piece of alternative evidence for convergent validity (Hair, Black, Babin, & Anderson, 2014, p. 619; Kline, 2016, p. 313).

Table 6. Cronbach's alpha, AVE, and CR

Factors	α	AVE	CR
1. Teacher Approachability	0.919	0.43	0.922
2. Positive Peer Relationships	0.839	0.38	0.845
3. Lack of Identity Safety	0.789	0.36	0.791

Further, to ensure discriminant validity the AVE results were compared with that of squared correlation estimates between the constructs. Table 7 shows that squared correlations of the constructs are smaller than AVE supporting discriminant validity of the scale (Hair et al., 2014). However, the AVE statistics are below the 0.50 rule of thumb as noted earlier.

Table 7. The square of the between-factor correlation estimates compared to AVE

Factors	1	2	3
1. Teacher Approachability	(0.43)		
2. Positive Peer Relationships	0.31**	(0.38)	
3. Lack of Identity Safety	0.12**	0.34**	(0.36)

Note: AVE statistics are given in parentheses, ***p* < 0.001

Therefore, as another measure for discriminant validity, the goodness-of-fit indices of the scale was computed for the one-factor model and then the results were compared to that of the three-factor model. The comparison between these fit indexes, the one and three construct models, indicated substantially different results. The results for the one-factor model, examined without modification, displayed a poor fit. Only the NNFI (0.90) and CFI (0.91) were within the acceptable ranges. However, the three-factor solution suggested a good-fitting model with NNFI and CFI of perfect indices as noted before. These findings suggest that 32 items in the scale represent three separate constructs rather than one, which is a good sign of discriminant validity according to Hair et al. (2014).

5. DISCUSSION and CONCLUSION

An attempt was made in the present study to develop and cross-validate a measurement scale as regards the student perception of a safe learning environment in the Turkish context by employing both PCA and PAF to compare results, but the priority was given to PCA as a conservative approach. Then, a CFA was run within a correlated traits model. Before performing the main analyses, the newly generated items (*N* = 85) underwent two rounds of expert review in terms of their relevance to the content domain and their comprehensibility to the target group. Item elimination or revision was carried out based on the CVI and modified kappa statistics at the initial stages of the study. These statistics yielded a strong content validity in terms of both the relevance and clarity of the remaining 59 items. The subsequent EFA, based on the data collected from 556 lower secondary students, suggested a scale with a nine-factor solution at the beginning. Nevertheless, three factors having at least 5 % explained variances were preferred, as was also indicated by the inflexion point in the scree plot.

However, to prevent over-estimation in factor determination, Horn's parallel analysis was used, where the real eigenvalues were compared to those of randomly generated ones (Patil et al., 2008, Williams et al., 2010). As a result, a three-factor solution with 32 items was supported. These factors, extracted through PCA using Promax rotation, accounted for 50.622% of variances in total, which is considered adequate in the field of "humanities" (Williams et al., 2010, p. 6). However, PAF produced relatively smaller explained total variances of 45.546 %. The remaining unexplained variance, however, could be related to the other influencing factors that might affect student perception of a psychologically safe learning environment in the classroom setting. This, however, might need further investigation to reveal its other dimensions.

Besides, the factors were labeled as Teacher Approachability, Positive Peer Relationships, and the Lack of Identity Safety according to the existing literature (Boostrom, 1998; Beamon, 2001; Holley & Steiner; 2005; Foldy et al., 2009). All the items under these factors had loadings of over 0.58 compared to the significant cutoff recommended (i.e. 0.45; Tabachnick and Fidell, 2013) and the item-total correlations were significant ($p < 0.01$). All of the items were in Likert type with five response categories, that is, 5 = strongly agree, 4 = agree, 3 = undecided, 2 = disagree, and 1 = strongly disagree. Besides, all the items belonging to the third factor, Lack of Identity Safety, are negative either by meaning or by form. Therefore, they must be reverse-coded in future use, as were done in this study before the analyses.

Furthermore, the first and second-order CFA validated the predicted construct of SLEPS by EFA. The factor loadings, *t*-test statistics, and goodness of fit indices indicated that the measurement model created, without modification, is within the acceptable standards to measure student perceptions as regards the safe learning climate in the classroom. All the factor loadings were ≥ 0.45 with significant *t* statistics ($p < 0.001$). Likewise, the Cronbach's alpha test of reliability after the EFA and CFA indicated that the variables in the sub-scales have an adequate level of internal consistency, although small differences were noticed in between. Besides, the total scores for the subscales can range from 16 to 80 for the first factor, 9 to 45 for the second, and 7-35 for the third. These scores could be interpreted or compared according to their arithmetic means. An increase in the mean of these scores may explain an increase in the students' perceived psychological safety in the classroom.

Furthermore, the convergent validity of the scale was supported according to the CR statistics. The comparison between the AVE statistics and the square of the correlation estimates between factors as well as the comparison between the fit indexes of one and three-factor models indicated good evidence for the discriminant validity. Simply put, the AVE for each construct was larger than its squared correlation and the 32 items in the scale support three construct model rather than one construct model. Taken together, in the wake of these results, this measurement model, the SLEPS (Safe Learning Environment Perception Scale), is a valid and reliable instrument to be used in future research.

Given that all the items in this measurement scale are in Turkish, only students who speak this language can respond it. In addition, the analyses were done based on the heterogeneous data from the lower secondary students (grade 5-8), studying at the public and private schools and the gifted center. This heterogeneity suggests the applicability of the scale in different contexts. The SLEPS, developed in the present study, could be utilized in collecting data from different schools or educational centers that accommodate students with diverse sociocultural and economic backgrounds. Although it was designed for the lower secondary students, its utility is assumed at the upper secondary level, too.

ORCID

Sayed Masood Haidari https://orcid.org/0000-0003-3221-6343
Fazilet Karakuş https://orcid.org/0000-0002-6455-9845

5. REFERENCES

Aldrich, J. O. & Cunningham, J. B. (2016). *Using IBM® SPSS Statistics* (2nd ed.). USA: SAGE Publications, Inc.

Beamon, G. W. (1993). Is Your Classroom "Safe" for Thinking? Introducing an Observation Instrument to Assess Classroom Climate and Teacher Questioning Strategies. *Research in Middle Level Education*, *17*(1), 91-110.

Beamon, G. W. (2001*). Making Classroom "Safe" for Adolescent Learning*. Paper Presented at the 53rd Annual Meeting of the American Association of Colleges for Teacher Education, Dallas, Texas.

Boostrom, R. (1998). 'Safe spaces': reflections on an educational metaphor. *Curriculum Studies*, *30*(4), 397-408).

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd ed). USA: The Guilford Press.

Capern, T. & Hammond, L. (2014). Establishing Positive Relationships with Secondary Gifted Students and Students with Emotional/Behavioural Disorders: Giving These Diverse Learners What They Need. *Australian Journal of Teacher Education, 39*, 46-67. doi:10.14221/ajte.2014v39n4.5

Çokluk, Ö. & Koçak, D. (2016) Using Horn's Parallel Analysis Method in Exploratory Factor Analysis for Determining the Number of Factors. *Educational Sciences: Theory and Practice*, *16*, 537-551. doi:10.12738/estp.2016.2.0328

Field, A. (2009). *Discovering Statistics Using SPSS* (3rd ed.). Oriental Press, Dubai: SAGE Publication.

Foldy, E. R., Rivard, P. and Buckley, T. R. (2009). Power, Safety, and Learning in Racially Diverse Groups. *Academy of Management Learning & Education*, *8*(1), 25–41.

Frisby, B. N., Berger, E., Burchett, M., Herovic, E. & Strawser, M. G. (2014). Participation Apprehensive Students: The Influence of Face Support and Instructor– Student Rapport on Classroom Participation. *Communication Education*, *63*, 105 - 123. doi:10.1080/03634523.2014.881516

Gayle, B. M., Cortez, D. & Preiss, R. W. (2013). Safe Spaces, Difficult Dialogues, and Critical Thinking. *International Journal for the Scholarship of Teaching and Learning*, *7*(2). doi:10.20429/ijsotl.2013.070205

Gillen, A., Wright, A. & Spink, L. (2011). Student perceptions of a positive climate for learning: a case study. *Educational Psychology in Practice*, *27*, 65 - 82. doi:10.1080/02667363.2011.549355

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate Data Analysis* (7th ed.). USA: Pearson Education Limited.

Holley, L. C. & Steiner, S. (2005) Safe Space: Student Perspectives on Classroom Environment, *Journal of Social Work Education*, *41*(1), 49-64.

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

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modelling* (4th ed). USA: The Guilford Press.

Lynn, M.R. (1986). Determination and Quantification of Content Validity. *Nursing Research, 35*, 382–385.

Patil, H. V., Singh, S. N., Mishra, S. & Donavan, D. T. (2007). *Parallel Analysis Engine to Aid in Determining Number of Factors to Retain* [Computer software], Available from http://ires.ku.edu/~smishra/parallelengine.htm

Patil, H. V., Singh, S. N., Mishra, S. & Donavan, D. T. (2008). Efficient theory development and factor retention criteria: Abandon the 'eigenvalue greater than one' criterion. *Journal of Business Research*, *61*, 162-170. doi:10.1016/j.jbusres.2007.05.008

Pilot, D. F. and Beck, C. T. (2006). The Content Validity Index: Are You Sure You Know What's Being Reported? Critique and Recommendations. *Research in Nursing and Health*, *29*, 489-497. doi:10.1002/nur.20147

Pilot, D. F., Beck, C. T. and Owen, S. V. (2007). Focus on Research Methods: Is the CVI an Acceptable Indicator of Content Validity? Appraisal and Recommendations. *Research in Nursing and Health, 30,* 459-467. doi:10.1002/nur.20199

Pituch, K. A. & Stevens, J. P. (2016). *Applied Multivariate Statistics for the Social Science* (6th ed). New York: Routledge.

Raghallaigh, M. N. & Cunniffe, R. (2013). Creating a safe climate for active learning and student engagement: an example from an introductory social work module. *Teaching in Higher Education*, *18,* 93-105. doi:10.1080/13562517.2012.694103

Seçer, I. (2013). *SPSS ve LISREL ile Pratik Veri Analizi: Analiz ve Raporlaştırma [Practical Data Analysis with SPSS and LISREL: Analysing and Reporting]*. Ankara: Anı Yayıncılık.

Steven, J. P. (2009). *Applied Multivariate Statistics for the Social Sciences* (5th ed). USA: Routledge.

Tabachnick, B.G. & Fidell, L. S. (2013). *Using Multivariate Statistics (6th Edition).* USA: Pearson Education, Inc.

Turner, S. & Braine, M. (2015). Unravelling the 'Safe' concept in teaching: what can we learn from teachers' understanding?. *Pastoral Care in Education, 33*(1), 47-62.

Walts, C. F., Strickland, O. L., & Lenz, E. R. (2010). *Measurement in Nursing and Health Research* (4th ed). New York: Springer Publishing Company, LLC.

Wanless, S. B. (2016). The Role of Psychological Safety in Human Development. *Research in Human Development*, *13,* 6-14. doi:10.1080/15427609.2016.1141283

Williams, B., Onsman, A., and Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Journal of Emergency Primary Health Care (JEPHC), 8*(3), 1-13.

Zamanzadeh, V., Ghahramanian, A., Rassouli, M. Abbaszadeh, A. Alavi-Majid, H. and Nikanfar, A. R. (2015). Design and Implementation Content Validity Study: Development of an instrument for measuring Patient-Centered Communication. *Journal of Caring Sciences, 4,* 165-178. doi:10.15171/jcs.2015.017

APPENDIX

The SLEPS

Öğretmen Yaklaşımı (Teacher Approachability)

36) Sınıfta öğretmenlerime güvenirim.
I trust my teachers in the class.

45) Sınıfımda öğretmenler hepimize eşit davranır.
Teachers treat all of us equally in my class.

53) Sınıfımda öğretmenler ayrımcılık yapmaz.
Teachers do not discriminate in my class.

6) Öğretmenlerimiz derslere aktif olarak katılmamız için çaba gösterir.
Our teachers make every effort to encourage our active participation in the lessons.

30) Öğretmenlerim öğrenme eksikliklerimle ilgili sorunlarımı çözmeye çalışır.
My teachers try to solve my problems in learning.

24) Öğretmenlerimiz bize iyi davranarak sınıfta güvende olduğumuzu hissettirir.
Our teachers make us feel safe in the classroom by treating us well.

34) Öğretmenlerim arkadaşlarımızla olan sorunlarımızı çözmemize yardım eder.
My teachers help us to solve our problems with our friends.

40) Öğretmenlerimiz kendimizi özgürce ifade edebileceğimiz ortamlar yaratır.
Our teachers create environments where we can express ourselves freely.

58) Öğretmenlerim sınıfta eğlenceli bir şekilde ders anlatır.
My teachers teach in an enjoyable way in the class.

48) Sınıfımdaki etkinliklerde herkese eşit katılım imkânı sağlanır.
Everyone gets equal opportunities to participate in the activities in my class.

28) Derste verilen görevlerde-etkinliklerde hata yaptığımda öğretmenlerim kızmaz.
My teachers do not get angry if I make mistake in the tasks given in the class.

59) Sınıfımda öğretmenler hepimize arkadaşça bir tavır sergiler.
Teachers are friendly to all of us in the class.

18) Sınıfta öğretmenlerim beni sabırla dinler.
My teachers listen to me patiently in the class.

47) Sınıfta öğretmenlerimle rahatlıkla iletişim kurarım.
I can easily communicate with my teachers in the class.

21) Sınıfta öğretmenlerim beni başka öğrencilerle kıyaslamaz.
My teachers do not compare me to other students in the class.

Pozitif Akran İlişkileri (Positive Peer Relationships)

19) Sınıfımda herkes birbirine iyi davranır.
Everyone treats each other well in my class.

51) Sınıf arkadaşlarım beni önemser.
My classmates care about me.

42) Sınıf arkadaşlarım yeni düşünceleri hoş karşılarlar.
My classmates welcome new ideas.

8) Sınıf arkadaşlarımla iyi anlaşırım.
I get along well with my classmates.

3) Sınıf arkadaşlarım düşüncelerime saygı duyar.
My classmates respect my ideas.

33) Sınıfımda fikir ayrılıkları kavgaya neden olmaz.
Disagreements in my class do not cause a fight.

41) Sınıfımda zıt düşünceler rahatlıkla paylaşılır.
Opposite ideas are easily shared in my class.

31) Sınıfta düşüncelerimi çekinmeden paylaşırım.
I share my thoughts without hesitation in the class.

22) Sınıfımda arkadaş ayrımı yapılmadan grup çalışmaları yürütülür.
The group works in my class are conducted without discrimination between friends.

Kimliksel Güven Eksikliği (Lack of Identity Safety)

55) Sınıfımda soru sorduğumda sınıf arkadaşlarımın dalga geçeceklerini düşünürüm.
I think my classmates will make fun of me when I ask questions in my class.

39) Sınıfımda hakarete maruz kalmaktan korkarım.
I am afraid of being humiliated in my class.

50) Sınıf ortamında söz almaktan çekinirim.
I am afraid to speak in the class.

44) Sınıfımda fikirlerimin yanlış anlaşılmasından endişe ederim.
I fear that my ideas may cause misunderstooding in my class.

57) Sınıfımda konuşmak istediğimde bana söz hakkı verilmez.
I am not given the right to speak in my class when I want to.

32) Sınıfımda öğrenme sırasında yanlış yaptığımda dalga geçilmekten korkarım.
I am afraid of being made fun of when I make a mistake in my class.

46) Sınıfta çok soru sorduğumda arkadaşlarım olumsuz tepki gösterir.
My classmates show negative reactions when I ask many questions in the class.

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 461–475
https://dx.doi.org/10.21449/ijate.561154

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

Measuring Nature of Science Views of Middle School Students

Yalçın Yalaki [1,*], Nuri Doğan[2], Serhat İrez[3], Nihal Doğan[4], Gültekin Çakmakçı[5], Başak Erdem Kara[6]

[1]Hacettepe University, Department of Primary Education, Ankara, Turkey
[2]Hacettepe University, Department of Educational Sciences, Ankara, Turkey
[3]Marmara University, Department of Mathematics and Science Education, Istanbul, Turkey
[4]Bolu Abant Izzet Baysal University, Department of Mathematics and Science Education, Bolu, Turkey
[5]Hacettepe University, Department of Mathematics and Science Education, Ankara, Turkey
[6]Hacettepe University, Department of Educational Sciences, Ankara, Turkey

ARTICLE HISTORY

Received: 07 May 2019
Revised: 03 August 2019
Accepted: 22 August 2019

KEYWORDS

Nature of science,
Middle school,
Assessment,
Multiple choice

Abstract: Developing scientific literacy for all students is the most often stated purpose of contemporary science education. Nature of science (NOS) is seen as an important component of scientific literacy. There are various perceptions of NOS in the science education community and NOS itself is an ever-changing construct. This makes it challenging to develop instruments for measuring understanding of NOS at different levels. Many instruments have been developed and are being developed to assess NOS learning, which indicates the importance attributed to this subject. In this study, we developed a multiple-choice test to measure NOS understanding of middle school students. The instrument was applied to 1397 middle school students. The 24 item multiple-choice test had KR-20 reliability coefficient of 0.74. A 12 item multiple-choice test created as a subset of the 24 items of the original test. This test was easier and had higher discrimination, which can provide useful measurement data about students' understanding of NOS for diagnostic or formative purposes.

1. INTRODUCTION

Developing scientific literacy for all students is a frequently stated purpose of contemporary science education curricula in many countries around the world. Nature of science (NOS) is seen as an important component of scientific literacy, and the importance of teaching NOS is emphasized by important educational policy documents and by many scholars (American Association for the Advancement of Science, 1990, 1993; Lederman, 1992; Matthews, 1998; Next Generation Science Standards Lead States, 2013; National Research Council, 1996). The importance given to the teaching and learning of NOS has increased steadily over the last 40 years and so the discussions about what NOS is and how to teach it (Lederman, 2007). Accordingly, there are different perceptions of NOS in the science education community and NOS itself has been an ever-changing construct (Matthews, 1998; Lederman, 2007; Abd-El_Khalick, 2014). This creates a problem for teaching and learning of NOS, as to teach

CONTACT: Yalçın Yalaki ✉ yyalaki@hacettepe.edu.tr Hacettepe Universitesi, Eğitim Fakültesi, Temel Eğitim Bölümü, 06800 Beytepe, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

something, a perception of it must be acknowledged so that teaching practices can be planned accordingly. The same problem also affects the assessment of NOS. To develop assessment instruments for NOS learning, a certain perception of NOS must be agreed.

Among alternative perceptions of NOS that are discussed in science education, the *consensus view*, *features of science view* and *family resemblance view* can be mentioned here as examples (Lederman, 1992; Matthews, 1994, Nola & Irzik, 2010). The consensus view argues that there is sufficient consensus on certain tenets of science that should be taught in schools. The features of science view argue that the list of consensus view tenets is arbitrary and there may be more than one list of consensus view tenets or the list can be extended. The family resemblance view argues that a more comprehensive view of science can be established based on similarities among different science fields. However, as Abd-El_Khalick (2014) says, "…the construct (or constructs) in currency in the field of science education is the NOS construct or are those constructs being assessed (p. 628)." One of the most popular and most assessed construct of NOS is the so-called consensus view construct. Lederman (2007) describes the seven tenets of NOS, which constitute the consensus view as:

1. Scientific knowledge is based on evidence: science is based on direct or indirect observation of the natural world. Science is not only based on empirical evidence, it is also based on logical inferences related to evidence. Scientific knowledge is supported through experimental data but it is never proved. Observation and inference should not be confused with each other. Scientists may have different inferences of same observations.
2. Scientific knowledge is durable but also tentative: scientific knowledge is stable but it is never certain or unequivocally true. Scientific knowledge changes through evolutionary and revolutionary processes. Scientific knowledge may change with new data or reevaluation of existing data.
3. Scientific knowledge involves subjectivity: Scientists' prior knowledge, experience, values, beliefs, education and expectations influence their study and the conclusions they reach. As a field of science matures, the level and amount of disagreements among scientists may decrease.
4. Scientific knowledge involves creativity and imagination: Scientists use their creativity and imagination in every stage of their scientific work. Creativity and imagination is an important factor that differentiates scientists from one another.
5. Science is a social activity thus it is influenced by the sociocultural environment: Political establishment, social values, economic conditions, and cultural structure influence how, what and to what degree scientists study a subject and how they apply their findings.
6. Scientific theories and laws: Theories are scientific explanations while laws are scientific descriptions of the natural phenomenon. They serve different purposes in science and there is no hierarchical relationship between them.
7. Scientific method: There is no one universal scientific method that all scientists follow that guarantees scientific discovery. Many different fields of science use many different methods to produce scientific knowledge.

The availability of various assessment instruments that are designed based on the consensus view of NOS is among the reasons for the popularity of this view in the field of science education.

Abd-El_Khalick (2014) provides a detailed landscape of NOS assessment instruments of the past 60 years. His analysis shows a trend of shifting towards open ended instruments from forced choice instruments (which include multiple-choice, Likert and agree-disagree type of instruments). In his analysis of the literature, Abd-El-Khalick shows that three NOS assessment instruments Test on Understanding Science (TOUS) (Cooley & Klopfer, 1961); Views on

Science-Technology-Society (VOSTS) (Aikenhead & Ryan, 1992); and Views of Nature of Science (VNOS) including its alternative forms (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002) dominated all the NOS assessment instruments developed in the last 60 years. He argues that these instruments collectively constituted more than 50% of the used instruments in published research in this period. Of these instruments, TOUS was a theoretically developed forced-choice instrument (meaning its items were developed from a theoretical perspective of NOS), VOSTS was an empirically developed forced-choice instrument (meaning its items were developed based on empirical data) and VNOS was an open ended instrument with corresponding interviews. This trend supports author's claim that there is a shift towards open ended instruments for NOS assessment in recent years. Abd-El-Khalick also argues that the VNOS instruments are the most popular in the field in recent years.

The forced choice instruments of NOS assessment are criticized for their shortcomings as open-ended instruments became more prevalent (Abd El-Khalick, 2014; Aikenhead, 1988; Lederman et al., 2002; Lederman & O'Malley, 1990). The stated shortcomings of forced-choice instruments can be summarized as:

- Assumption that respondents understand the forced choice item the same way as developers.
- Validity of these instruments is threatened because of difficulties in interpreting respondents' choices.
- These instruments often embody a specific theoretical model of NOS which reflect developers' philosophical positions and preferences which are imposed on respondents by the choices provided.
- Respondents' NOS views are often fragmented and lacking, which makes it difficult to capture their views through forced choice instruments.
- Likert scale instruments are particularly problematic because they generate higher levels of ambiguity.

These arguments are fair, although some of them are also valid for open ended questionnaires. For one, every questionnaire, whether open ended or forced choice, is designed with a philosophical position in mind. It would be quite difficult to create a value free instrument. Secondly, open ended questionnaires do not gurantee clearer and less fragmented responses. We can argue out of experience that it can be very challenging to interpret written answers to open ended qestions, especially if they are short or unclear. Thirdly, respondends can still understand open ended questions very differently than the developers intended similar to the forced choice questions. Another problem with the open ended questionnaires is the scorer bias, which can be an important problem if necessary precautions are not taken. On the other hand, despite the above shortcomings, forced choice instruments have some major advantages. These advantages include scalability, ease of administration, ease of scoring, and ease of data analysis. Of course these benefits do not excuse the above shortcomings, but we believe that a relatively short multiple-choice test developed based on common respondent views about NOS that are reported in the literature rather than a theoretical view of NOS could still be useful for diagnosis of student views while keeping in mind its limitations.

Parallel to the developments around the world about NOS teaching and learning and assessment of NOS views, the last two primary school science curricula prepared by the Turkish Ministry of National Education (MEB, 2013; 2018) emphasize NOS as a component of scientific literacy. To contribute to the purpose of achieving scientific literacy in schools in Turkey, a long-term professional development program for science teachers about teaching and learning of NOS was organized as a research project and it was implemented with funding from The Scientific and Technological Research Council of Turkey (TUBITAK). Our project was titled "Continuing Teacher Professional Development to Support the Teaching about Nature of

Science (BIDOMEG)" which was carried out in cooperation with Abant Izzet Baysal University, Hacettepe University, Marmara University, Ministry of Education General Directorate of Teacher Training and Development, and Bolu Provincial National Education Directorate. In this article, we reported about the ScienTest study, a multiple-choice test and its development process which was aimed at measuring the NOS views of middle school students.

For the main data collection in the study, we used the VNOS-D instrument (Lederman & Khishfe, 2002), to measure seventh grade middle school students' views on NOS. The targeted NOS themes in the instrument were: 1- scientific knowledge is based on empirical evidence, 2-observation and inference are different from one other, 3- scientific knowledge is reliable but open to change, 4- creativity and imagination play an important role in the emergence of scientific knowledge, 5- scientists can be subjective during scientific studies, 6-scientific models are abstract and approximate versions of the reality. The sixth theme was added to the instrument as an extra dimension. The VNOS-D instrument is an open ended instrument and requires written answers. The written answers then need to be coded by various scorers. In large scale applications, the application of this instrument bears many difficulties. The open ended questions require young students to express their opinions in writing, which is often challenging and most of them prefer to write short answers as we very often observed during the study. Also coding written answers, especially short answers, can be challenging as it is often not clear which category the answer falls into. With close to 1400 participants, conducting interviews were not practical to clarify students' ideas. In addition, scorer errors, which can occur when the test items are evaluated and coded by different individuals, can affect the reliability of the instrument. Given the difficulties of implementing the VNOS-D in large scale, we decided to develop a multiple-choice instrument that measured the same six themes which can be applied from fifth to eighth grade levels. Development of this instrument was not a planned outcome of the project from the beginning, but rather the idea of developing such an instrument appeared with the challenges of large scale measurement.

2. METHOD

A 24 item test for the six NOS themes mentioned above was developed (see Appendix 1). Table 1 shows the targeted NOS themes and the corresponding items that were designed to measure student views about these themes. The items had three choices, each representing a different NOS understanding. One of these options represented an understanding at the targeted level (informed level in VNOS terms). The other choices were selected based on alternative conceptions of NOS that were reported in the literature. In order to confirm the validity of the test, four experts (the authors) have reviewed items and proposed changes, which were implemented.

Table 1. Targeted NOS themes and corresponding items in ScienTest

Items		Related NOS theme
1 2 3	13 14 15	Science is based on empirical evidence.
4 5	16 17	Scientific knowledge is tentative.
6 12	18 24	Scientific knowledge involves creativity and imagination.
7 8	19 20	Scientific knowledge involves observation and inference.
9 10	21 22	Scientific knowledge involves subjectivity.
11	23	Scientific models do not reflect exact reality.

2.1. Study Group

For the pilot study of ScienTest, two different forms of the instrument were prepared at the beginning of the study in an effort to find out which form is more reliable and these forms were applied to a total of 183 middle school students. One of the forms had 12 items and six choices and students were asked to mark all of the choices they preferred. The other form had 24 items and three choices, one of them being the desired choice. The analysis of data showed that the three-choice 24 item form was more reliable than the 12 item form (KR-20 0.641 vs 0.615). After the item analysis and validity assessments were made on the 24 item multiple-choice form, some items were revised and the final form was applied to 1397 middle school students in the spring semester of 2013. The reliability coefficient of the test (KR-20 value) was determined to be 0.740 in this large-scale application. The detailed data analysis is explained below.

2.2. Data Analysis

ScienTest's item and test parameter estimates were made according to both Classical Test Theory (CTT) and Item Response Theory (IRT). The analysis results were cross-checked based on two theories. In the analysis of data, TAP 14.7.4 software was used for analysis based on the Classical Test Theory, and IRTPRO software was used for the analysis based on the Item Response Theory. Parallel analysis based on the tetrachoric correlation matrix for factor analysis was performed with FACTOR 10.6.01 software program. Finally, Confirmatory Factor Analysis was performed with MPlus7 software program.

Data analysis took place in several stages. Firstly, exploratory factor analysis was carried out with Parallel Analysis method for the 24-item ScienTest to examine the structure of the data. After factor analysis, model-data fit of the IRT models (Rasch, 2PL and 3PL) were examined. The -2Log-Likelihood values were used to examine the fit of the IRT models' fit with data and the -2Log-Likelihood value differences for each model were compared with the chi-square difference test. Afterwards, CTT and IRT analysis were made. At the last stage, the best performing items in the two halves of ScienTest were selected, taking into account the fact that each dimension would be measured, and a 12-item final test was established and a confirmatory factor analysis was applied on this test.

3. RESULT / FINDINGS

In this section, the results of factor analysis, model-data fit, item and test analysis based on CTT and IRT, and confirmatory factor analysis of the final test are presented.

3.1. Factor Analysis

As a result of the Parallel Analysis based on the tetrachoric correlation matrix applied to the 24-item ScienTest to determine the test structure, when the values of KMO and Bartlett were analyzed it was found that the data structure was suitable for factor analysis (KMO = 0.845, Bartlett's test of sphericity $\chi2$ = 2957.5, p = 0.00010). The results of the factor analysis are presented in Table 2.

The eigenvalue for the first factor is approximately 3.2 times the eigenvalue for the second factor. Parallel analysis result also suggests a one-factor structure. According to this result, it is accepted that the data has a one-factor structure. One factor explains the 22.3% of the variance of test scores. Factor loadings were found to be in the range of 0.102 and 0.668. Büyüköztürk (2012) suggested that factor loadings should be at least .30. Items 2, 13, 16 and 17 have factor loadings below 0.30.

Table 2. Factor analysis results for ScienTest

Item	*Factor loadings*	*Item*	*Factor loadings*
1	0.418	13	0.150
2	0.183	14	0.592
3	0.376	15	0.653
4	0.400	16	0.102
5	0.459	17	0.260
6	0.556	18	0.668
7	0.455	19	0.512
8	0.382	20	0.474
9	0.400	21	0.587
10	0.499	22	0.445
11	0.440	23	0.300
12	0.613	24	0.369
Explained variance ratio	%22.3		

3.2. Model-Data Fit

Three different IRT models (Rasch, 2-Parameter Logistics and 3-Parameter Logistics) were used for data analysis based on IRT. The -2Log-Likelihood, AIC and BIC values were used to determine which model has the best fit on ScienTest data. The model with the smallest AIC and BIC values is interpreted as the best model (Wang & Liu, 2005). The obtained values were presented in Table 3.

Table 3. Model-data fit indexes

	-2Log-Likelihood	*AIC*	*BIC*
Rasch Model	41901.37	41949.37	42075.18
2 PLM	41400.89	41496.89	41748.51
3 PLM	41141.70	41285.70	41663.13

The model with the lowest Log-Likelihood, AIC and BIC values is the 3 parameter logistic model (3 PLM). In addition, for each model, the difference of -2loglikelihood values were compared with the chi-square difference test to investigate model fit. At this point, the $\chi 2$ value on the $\chi 2$ table was found first ($\chi 2_{(24,\ 0.05)} = 36.415$) and the value of -2Log Likelihood was compared with that $\chi 2$.

- 1PLM-2PLM: $\chi 2 = (\text{-2Log-Likelihood}_{1PLM}) - (\text{-2Log-Likelihood}_{2PLM}) = 500.48 > 36.42$, the 2-Parameter Logistic Model (2 PLM) is more significant than the Rasch model, that is, 2PLM shows better fit with the data.
- 2PLM-3PLM: $\chi 2 = (\text{-2Log-Likelihood}_{2PLM}) - (\text{-2Log-Likelihood}_{3PLM}) = 259.19 > 36.42$, the 3-Parameter Model provides a better fit than the 2-Parameter Model. The fact that this value is above the critical value indicates that the analysis of data with 3 PLM will make a significant difference.

As a result of the analysis, it is found that the best fitting model with data is 3-Parameter Model.

3.3. Descriptive Statistics

The total number of questions in the test was 24 and the number of respondents was 1397. Table 4 presents the descriptive statistics calculated for the ScienTest in the framework of CTT. The highest score that can be taken from this test is 24, where the correct answers are marked as '1' and the wrong answers are marked as '0'. The reliability coefficient calculated with the KR-20 method is .740. The reliability calculated by the Sperman-Brown split-half method (odd-even) was calculated as 0.763 and the McDonald's Omega was calculated as 0.83. McDonald's Omega is a reliability coefficient used for congeneric measurements, which is defined as measurements that items' factor loadings are different (McDonalds, 1985). When the factor loadings in Table 2 are examined, it can be said that the factor loadings differ, so the congeneric measurement can be applied. In this case, the appropriate reliability coefficient to use was McDonald's Omega, which had a value of 0.83. This value indicates that the test is reliable at an acceptable level (> .70) (Nunnaly, 1973). The test scores' mean was calculated as 13.081 (standard deviation = 4.355). The average difficulty value of the test was 0.545, while the average discrimination value was 0.378. When the values of the skewness and kurtosis were examined, it was found that they varied between -2 and +2; this is considered to be a sign of the normal distribution test scores (Pallant, 2005).

Table 4. Descriptive statistics for ScienTest

Sample size	1397	Kurtosis	-0.538
Mean	13.081	Reliability	0.740 (KR-20)
Average Difficulty	0.545		0.763 (Split Half)
Average Discrimination	0.378		0.83 (McDonald's Omega)
Median	13.00	Variance	18.965
Skewness	0.160	Standard Error	2.22

3.4. Item Analysis

The item difficulty and discrimination parameters for 24 items in the test are presented in Table 5. The item discrimination given in the framework of CTT has been interpreted using the point biserial correlation value. In the context of IRT analysis, 'a' parameter means item discrimination, 'b' parameter means item difficulty, and 'c' parameter is interpreted as guessing parameter.

When the analysis results in Table 5 are examined, it is seen that the item difficulties according to the CTT are between 0.26 and 0.74, and the average difficulty of the test is 0.545. This value can be regarded as an indicator that the test is at medium difficulty (Haladyna, 2004). The item discrimination values are found to be in the range of 0.17 and 0.44. Ebel (1965) stated that items with discrimination values smaller than 0.20 should be thrown away or completely replaced, while items with discrimination values between 0.20-0.29 should be corrected. Items with a discrimination value of .30 and above have a sufficient level of discrimination. According to this criterion, items numbered 2, 13 and 16 should be reexamined.

The IRT item analysis was interpreted only considering the 3 PLM estimates. The values in Table 5 show that the value of "a", which represents item discrimination, varies between 0.38 and 5.32. The "b" values, representing item difficulty, were in the range of -1.09 to 3.40; the "c" parameter, which is the indicator of guessing, varied between 0.10 and 0.36. According to the difficulty parameter, the items are spread over a wide range; it can be said that the test contains questions from all levels. Given the average difficulty parameter (b_{mean} = 0.51), it was found that the test's average difficulty was above the mean (b> 0). The estimated average guessing parameter was 0.23.

Hambleton and Swaminathan (2010) recommend paying attention to the items in cases that standard error value of item parameters exceed 1. When the standard error values for the parameters are examined, the standard error value (sh_{13} = 9.38, sh_{16} = 9.38) for parameter "b" for item 13 and parameter "a" value for item 16 is found to be higher than 1. According to the results of the IRT analysis, items 13 and 16 should be revised.

Table 5. Item analysis results based on CTT and IRT

	CTT		IRT					
			1 PL	2 PL		3 PL		
	Difficulty	Discrimination	b	a	b	a	b	c
1	0.55	0.41	-0.23	0.74	-0.30	1.73	0.78	0.36
2	0.26	0.23	1.21	0.33	3.24	2.96	1.85	0.21
3	0.74	0.31	-1.20	0.66	-1.73	0.72	-1.09	0.21
4	0.61	0.37	-0.52	0.72	-0.69	0.86	0.05	0.20
5	0.7	0.38	-0.96	0.87	-1.10	0.97	-0.64	0.18
6	0.68	0.44	-0.88	1.18	-0.82	1.66	-0.18	0.28
7	0.51	0.41	-0.05	0.75	-0.06	1.07	0.60	0.22
8	0.45	0.37	0.24	0.64	0.36	2.43	1.18	0.33
9	0.55	0.37	-0.23	0.70	-0.31	0.92	0.35	0.20
10	0.67	0.42	-0.82	0.94	-0.89	1.02	-0.53	0.15
11	0.58	0.39	-0.40	0.78	-0.50	1.10	0.23	0.24
12	0.66	0.48	-0.76	1.34	-0.67	1.63	-0.29	0.19
13	0.4	0.21	0.47	0.23	1.83	0.32	3.40	0.19
14	0.72	0.43	-1.08	1.23	-0.99	1.29	-0.79	0.11
15	0.7	0.49	-0.97	1.53	-0.79	1.66	-0.61	0.10
16	0.35	0.17	0.71	0.17	3.60	2.62	2.26	0.33
17	0.31	0.29	0.91	0.45	1.84	1.65	1.87	0.23
18	0.69	0.5	-0.94	1.65	-0.74	2.11	-0.38	0.20
19	0.53	0.44	-0.13	0.96	-0.15	1.36	0.37	0.20
20	0.49	0.41	0.06	0.84	0.07	1.73	0.74	0.26
21	0.57	0.47	-0.33	1.20	-0.31	1.75	0.16	0.21
22	0.55	0.4	-0.22	0.80	-0.27	1.90	0.70	0.34
23	0.35	0.33	0.70	0.53	1.22	5.38	1.29	0.27
24	0.5	0.36	-0.02	0.63	-0.03	1.62	1.00	0.34
AVERAGE	*0.545*	*0.378*				*1.69*	*0.51*	*0.23*

In the ScienTest, the first 12 questions and the second 12 questions are designed to measure the same attributes. In other words, it can be said that the first 12 items and the next 12 items were designed as a parallel test. In addition to the analysis and results obtained on the 24-item form of the test, the detailed analysis made on the 12-item two halves may be more informative. Since the two halves of the test measure the same attributes, two parallel tests were analyzed with independent exploratory factor analysis (Parallel Analysis based on the Tetrachoric Correlation Matrix) and the findings are presented in Table 6.

First, the results of the factor analysis for the test consisting of the first 12 questions were examined. When KMO and Bartlett values were examined, it was found that the data structure was appropriate for factor analysis (KMO = 0.764, Bartlett's test of sphericity $\chi 2$ (66) = 971.3, p = 0.000010). When we look at the eigenvalues, it is seen that the eigenvalues of three factors have a value higher than 1, but the eigenvalue of the first factor is about 2.6 times the eigenvalue of the second factor. Parallel analysis result also suggests a one-factor structure. The factor loadings for all items load between 0.201 and 0.588 on the first factor, which supports the one-factor structure conclusion. On the other hand, according to the assumption that 12 items are collected in one factor, the explained variance ratio was calculated as 26.2%. When Table 6 is examined, the factor loadings of only 2 items from the first 12 items are below the critical value of .30. According to Reckase (1979), explained variance ratio of 20% is enough.

When the results of the factor analysis for the test consisting of the last 12 questions were examined, the values of KMO and Bartlett were found to be at the desired levels (KMO = 0.747, Bartlett's test of sphericity $\chi 2$ (66) = 1171.1, p = 0.000010). According to these results, it can be said that the data is suitable for the factor analysis. According to the results of factor analysis for the last 12 questions, there are three factors with an eigenvalue greater than 1. However, the eigenvalue of the first factor is about 2.6 times the eigenvalue of the second factor, and the number of dimensions proposed by the Parallel Analysis method is also 1. In this case, it was decided that the structure has one factor. The variance ratio explained by one factor was 27.1% and the factor loadings related to 12 items were varied between 0.094 and 0.648. The factor loadings of the items corresponding to items 13, 16, 17 and 23 on the whole test were below .30.

Table 6. Factor analysis results for two half tests

Items	*First 12 items*	*Last 12 items*
	Factor Loads	Factor Loads
1	0.469	0.144
2	0.201	0.648
3	0.378	0.738
4	0.454	0.094
5	0.489	0.269
6	0.545	0.701
7	0.487	0.533
8	0.470	0.449
9	0.421	0.627
10	0.510	0.455
11	0.459	0.266
12	0.588	0.360
Explained variance ratio	%26.2	%27.1

When we look at Table 2, it is noted that the factor loadings of items 2, 13, 16 and 17 are low (<.30) in the factor analysis results obtained from the whole test of 24 items. In addition, according to the results of the CTT and IRT analysis, items 2, 13 and 16 need to be re-examined. The results of the exploratory factor analysis of the two 12-question half tests were also examined and it was seen that the factor loadings of the items 2, 13, 16, 17 and 23 were low. Following a concerted evaluation of all the results obtained, it was decided to form a 12-item final test consisting of the best-performing items in the two halves to investigate each behavior.

3.5. Creating a 12 Item Sub-Test

The selected items in the final test are 1, 4, 5, 8, 10, 11, 12, 14, 15, 18, 19 and 21. Factor loadings and item statistics have been considered in the selection of these items. Descriptive statistics of the 12 items selected are presented in Table 7. The total number of items in the final sub-test is 12 and the number of respondents is 1397. When the descriptive statistics of the final sub-test are examined, it is seen that the average difficulty is .617 and the discrimination is .475. The average difficulty and discrimination values of the final sub-test are both higher than the initial 24-item test. In this case, the interpretation can be made that the sub-test became easier and the test discrimination became higher. The reliability coefficient calculated according to each method is lower than the 24-item test. It is normal to encounter this situation when it is thought that the value of reliability is affected by the number of items. If the acceptable level of reliability is considered to be .70 and above, it can be said that the sub-test is reliable at acceptable levels. The skewness and kurtosis values are in the range of -2 to +2 and it is accepted that test scores’ distribution is normal.

Before beginning confirmatory factor analysis, Mardia Test was conducted to check whether multivariate normality assumption was satisfied or not and it was seen that multivariate normality was not achieved (p <.05). Given that the structure of the data is categorical and not normally distributed, WLSMV and ULSMV, which are recommended estimation methods for this data (Brown, 2015), have been preferred. The results are given in Table 8.

Table 7. Descriptive statistics of the 12-item sub-test

Sample size	1397	Reliability	0.685 (KR-20)
Average difficulty	0.617		0. 716 (Split-Half)
Average Discrimination	0.475		0.798 (McDonald's Omega)
Median	8	Variance	7,409
Skewness	-0,282	Standard Error	1,528
Kurtosis	-0.685		

Table 8. Confirmatory factor analysis results for WLSMV and ULSMV techniques of final sub-test

WLSMV	df	χ2	χ2/df	RMSEA	CFI	TLI
	54	144.739*	2.68	0.035	0.961	0.952
ULSMV	df	χ2	χ2/df	RMSEA	CFI	TLI
	54	139.507*	2.58	0.034	0.960	0.952

* p < .001

At first χ2 test results were investigated among the confirmatory factor analysis results. The significance of p (p <.05) for this test is an indication that the model fit is weak. However, the chi-square statistic is a statistic that is highly influenced by the sample size. For this reason, the use of chi-square/df in large samples is recommended. If this value is between 3 and 5, acceptable fit is shown, and if it is smaller than 3, it shows perfect fit (Hair, Black, Babin & Anderson, 2009; Kline, 2015; Tabachnick & Fidell, 2013). Approximation of the goodness of fit index values (CFI and TLI) to 1 can be regarded as an indication that the model fits well with the data. For index values, 0.90-0.95 is acceptable, and above 0.95 indicates a good fit. On the other hand, if the RMSEA values indices are 0, it is perfect, and if it approaches 0, it is a good model fit. If this value is less than .03, perfect fit is accepted, if it is in the range of .03-.08, it is considered as an acceptable fit indicator (Brown, 2015; Hair et al., 2009). When all the values in Table 7 are considered together, it can be said that the one-factor model shows very good fit with the data according to the both WLSMV and ULSMV estimation methods (χ2 / sd =2.68-2.58, RMSEA =0.035-0.034, CFI =0.961-0.960, TLI = .952).

4. DISCUSSION and CONCLUSION

NOS teaching and learning is a dynamic field of study and so is the assessment of NOS learning. Many instruments are developed and continue to be developed to assess the understanding of this important construct (Abd-El-Khalick, 2014). The fact that so many instruments are being developed to assess NOS learning indicates the importance attributed to this subject. One recent example is the Nature of Science Instrument (NOSI) developed by Hacıeminoğlu, Yılmaz-Tüzün, and Ertepınar (2014). This instrument is a 13 item three point Likert scale developed to assess sixth, seventh, and eighth grade elementary students' NOS views. It focuses on four NOS themes which are "the difference between observation and inferences, tentativeness of scientific knowledge, role of imagination and creativity in scientific knowledge, and dependence of scientific knowledge on empirical evidence." The authors conducted the reliability study of this instrument with 782 students. Another example is Nature of Science View Scale (NOSvs) developed by Temel, Şen, and Özcan (2018). This instrument was developed with participation of 565 prospective teachers from different fields. The instrument is a 36 item five point Likert scale. The authors report that the final instrument measured five subscales which were 'definition and limits of science; scientific method; theory-laden and subjective nature of science; sociocultural embeddedness of science; and tentative and empirical nature of science.' All of the subscales had Cronbach alpha values above 0.70.

The above instruments were developed for an older audience than that of the ScienTest instrument. In this study, we targeted a younger audience with a multiple-choice test rather than the Likert scale which may have a higher level of ambiguity, especially with younger people. We wanted students to choose a view among given choices rather than express their degree of agreement with a view. The data analysis showed that the 24 item multiple-choice version of the test has KR-20 reliability coefficient of 0.74. As the data analysis show, a 12 item multiple-choice test created as a subset of the 24 ScienTest items created a final test with better mean difficulty and discrimination values. This 12-item sub-test was easier and had higher discrimination. This version of the test has sufficient reliability, albeit lower than the original test, and it still measures all of the NOS themes in the test. The shorter version of the test can be used with relatively younger students as it involves less reading.

In conclusion, we believe that this test can be used to collect data bout middle school students' NOS views. Multiple-choice tests have many disadvantages and also many advantages. As no measurement tool is perfect, this instrument is also not perfect, but it can provide useful measurement data about students' understanding of science for diagnostic or formative purposes.

Acknowledgements

This study was conducted under the "Continuing Teacher Professional Development to Support the Teaching about Nature of Science (BIDOMEG)" project supported by the The Scientific and Technological Research Council of Turkey (TUBITAK) with project number 111K517.

Conflict of interest

The authors declare no conflict of interest.

ORCID

Yalçın Yalaki https://orcid.org/0000-0003-0939-4766

5. REFERENCES

American Association for the Advancement of Science. (1990). *Science for All Americans*. New York: Oxford University Press.

American Association for the Advancement of Science. (1993). *Project 2061: Benchmarks for science literacy*. New York: Oxford University Press.

Abd-El-Khalick, F. (2014). The evolving landscape related to assessment of nature of science. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of Research on Science Education, Volume II* (pp. 635-664). New York: Routledge.

Aikenhead, G. S. (1988). An analysis of four ways of assessing student beliefs about STS topics. *Journal of research in science teaching*, *25*(8), 607-629.

Aikenhead, G. S., & Ryan, A. G. (1992). The development of a new instrument: 'Views on Science—Technology—Society'(VOSTS). *Science education, 76*(5), 477-491.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research (2nd ed.)*. New York, NY: Guilford Press.

Büyüköztürk, Ş. (2012). *Sosyal bilimler için veri analizi el kitabı [Data analysis handbook for social sciences]*. Ankara: Pegem Akademi.

Cooley, W. W. & Klopfer, L. E. (1961). *TOUS: Test on understanding science*. Princeton, NJ: Education Testing Service.

Hacıeminoğlu, E., Yılmaz-Tüzün, Ö., & Ertepınar, H. (2014) Development and validation of nature of science instrument for elementary school students. *Education 3-13, 42*(3), 258-283

Hair, J. F., Black, W. C, Babin, B.J. & Anderson, R. E. (2009). *Multivariate data analysis* (7. ed.). Upper Saddle River, NJ: Pearson Education.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Swaminathan, H. (2010). *Item response theory: principles and applications*. Norwell, MA: Kluwer Nijhoff Publishing.

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Lederman, N.G. (1992). Students' and Teachers' Conceptions of the Nature of Science: A Review of the Research, *Journal of Research in Science Teaching, 29*(4), 331- 359.

Lederman, N.G. (2007). Nature of science: Past, present, and future. In S.K. Abell &N.G. Lederman (Eds.), *Handbook of research in science education* (pp. 831–880). Mahwah, NJ: Lawrence Erlbaum Associates.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. (2002). Views of nature of science questionnaire (VNOS): Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching, 39*(6), 497-521.

Lederman, J. S., & Khishfe, R. (2002) Views of nature of science, Form D. Unpublished paper: Illinois Institute of Technology, Chicago, IL.

Lederman, N. G., & O'Malley, M. (1990). Students' perceptions of tentativeness in science: Development, use, and sources of change. *Science Education*, *74*(2), 225-239.

Matthews, M. R. (1998). In defense of modest goals when teaching about the nature of science. *Journal of Research in Science Teaching 35*(2), 161–174.

Milli Eğitim Bakanlığı [MEB]. (2013). *İlköğretim kurumları (ilkokullar ve ortaokullar) fen bilimleri dersi (3, 4, 5, 6, 7 ve 8. sınıflar) öğretim programı. [Primary schools (elementary and middle) science lesson (3, 4, 5, 6, 7 and 8th grades) curriculum]*. Ankara: MEB

Milli Eğitim Bakanlığı [MEB]. (2018). *İlköğretim kurumları (ilkokullar ve ortaokullar) fen bilimleri dersi (3, 4, 5, 6, 7 ve 8. sınıflar) öğretim programı. [Primary schools (elementary and middle) science lesson (3, 4, 5, 6, 7 and 8th grades) curriculum]*. Ankara: MEB

Nunnally, J. C. (1973). Research strategies and measurement methods for investigating human development. In J. R. Nesselroade & H. W. Reese, *Life-span developmental psychology: Methodological issues*. Oxford, England: Academic Press.

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

Next Generation Science Standards Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

Pallant, J. (2005). *SPSS survival manual: a step by step guide to data analysis using SPSS*. Maidenhead: Open University Press/McGraw-Hill,

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207-230.

Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Upper Saddle River, NJ: Pearson Education.

Temel, S., Şen, Ş. & Özcan, Ö. (2018) The development of the nature of science view scale (NOSvs) at university level. *Research in Science & Technological Education, 36*(1), 55-68

Wang, Y. & Liu, Q. (2005). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of stock–recruitment relationships. *Fisheries Research*, *77*, 220–225.

Appendix 1. Final version of the ScienTest (English translation)

Turkish version can be downloaded at: http://www.bilimindogasi.hacettepe.edu.tr/Biltest.pdf

1. There are statements about science below. Circle the one you think is correct.

a) Science is about the knowledge we see in the science courses.
b) Science is the new technologies that are invented and developed.
c) Science never produces a hundred percent certain knowledge, but it produce valid and reliable knowledge.

2. Which of the following do you think is a scientific discipline that is based on real experiments and observations?

a) Ufology (investigates the unknown objects seen in the sky that are called UFOs)
b) Biology (investigates living things, their structure and behavior)
c) Turkish (investigates the rules, use, reading and writing of the Turkish language)

3. Which of the following do you think is about a scientific study?

a) Creating a computer model of how a star is formed based on available data
b) Designing a new automobile model
c) Searching on the internet to learn how influenza spreads

4. Which of the following statements about scientific knowledge is true?

a) Scientific knowledge is objective; it does not change from person to person.
b) Scientific knowledge (theories, laws, hypotheses, etc.) can change with new studies and data.
c) There is only one way of producing scientific knowledge and that is the scientific method.

5. Which of the following statements about the knowledge you learned in the science and technology courses do you agree?

a) The knowledge in the Science and Technology textbooks are obtained through years of research and are unlikely to change.
b) The fact that new inventions like tablet computers and smart phones happen shows that the knowledge we read in textbooks may change one day.
c) The knowledge in Science and Technology textbooks are reliable and valid but this does not mean that this knowledge will never change in the future.

6. Do you think scientists use their creativity and imagination when they do research and experiments?

a) Whether scientists use their creativity and imagination or not depends on their field of study.
b) Science does not change from person to person; therefore, scientific studies are not influenced by creativity and imagination.
c) Scientists use their creativity and imagination in scientific studies and that is why sometimes they arrive at different conclusions.

7. It is known that all matter is made up of atoms. However, atoms' internal structure is too small to be seen even with the most powerful electron microscopes. Which of the following statements do you agree about the knowledge that scientists obtained about atoms?

a) Since we cannot see atoms, all of the diagrams and models created about atoms may not be entirely correct.
b) Knowledge about atoms are obtained through studies that have been conducted for a long time and became certain in present-day.
c) Atoms' structure can only be understood if powerful enough microscopes can be made that show their internal structure in the future; otherwise we cannot know anything about atoms.

8. Dinosaurs have lived on earth for a long time and they disappeared 65 million years ago. T-Rex is one of the most predatory dinosaurs known. How do you think scientists know that dinosaurs like T-Rex really existed and how much can they be sure of how they looked?

a) They are sure of their existence and appearance thanks to fossils and bone fragments that they found.
b) They can combine bone fragments to guess the body shape of a dinosaur, but they cannot be sure of their real look.
c) As there are pictures, models, films and documentaries about dinosaurs, scientists are sure of what they have looked like.

9. Turkey is a country that experience earthquakes often. As a result of scientific studies, scientists think that there may be an earthquake in Istanbul region in the near future. However, they expressed different opinions about the time and intensity of such an earthquake. Even though scientists have the same information, why do you think they have different opinions about this issue?

a) They have different opinions because there is no valid theory about this subject.
b) They have different opinions because they have not come together and thoroughly discussed the issue.
c) They have different opinions because they have different backgrounds, experience, knowledge and means.

10. The relationship of technological developments such as cell phones with cancer is being discusses. Studies about this relationship provided conflicting results. Some experts report that extensive use of cell phones increase the risk of cancer, while others could not find a relationship between cell phones and cancer. What do you think is the reason fort these conflicting results?

a) These kinds of conflicts may appear in the beginning of research, but eventually they are definitely resolved.
b) Scientists' preferred methods, their inferences and judgements may be different which may lead to conflicting results.
c) Science is objective and these kinds of conflicts should not exist. So one of the studies must be wrong.

11. There are different models that you use in schools (for example a model that shows internal organs, a cell model, and a DNA model, etc.). Scientists also use models when they investigate the nature. How much do you think these models reflect reality?

a) These models help us understand science subjects, but they are not real, they are only simplified versions of reality.
b) Models of very complex systems may not reflect reality, but models of simple things reflect reality.
c) If a model is well prepared, it reflects the reality.

12. At which stage/stages of their research (for example planning, doing an experiment, analyzing data, interpreting data, reporting the results, etc.) do you think scientists use their creativity and imagination?

a) All stages of a research can be done in different ways and creativity and imagination can play a role in all of the stages.
b) Creativity and imagination may play a role when planning a research but other than that it is not important.
c) I don't think they use their creativity and imagination in any stage of their research.

13. Which of the following statements about science do you think is correct?

a) Science provides certain, accurate knowledge.
b) Science allows us to reach the reality as a result of many studies conducted.
c) Science is based on experiments, observations, and logical inferences based on them.

14. Which of the following is a scientific field that is based on experiments and observations?

a) Mathematics (investigates numbers, shapes, geometry, operations, functions, etc.)
b) Chemistry (investigates matter, properties of matter, and how matter changes)
c) History (investigates the past events, people, institutions, and their relationships)

15. Which of the following do you think is a scientific study?

a) Conducting a controlled experiment on subjects to find out the effect of a medicine on cancer
b) Solving a very hard mathematical problem
c) Preparing educational TV programs about scientific topics such as genetic engineering

16. Which of the following statement about scientific knowledge do you think is correct?

a) Scientific knowledge is type of knowledge that is proven to be certain by experiments.
b) The differentiating feature that separate scientific knowledge from other knowledge is its testability.
c) All scientific knowledge becomes law after being proven in time.

17. Which of the following statements about knowledge in science textbooks do you think is right?

a) Only proven knowledge enters science textbooks, unproven knowledge cannot be in these books.
b) Some knowledge in textbooks may change in the future, but knowledge that became law never change.
c) All of the information in the science textbooks can possibly change in the future.

18. Do you think scientists use their creativity and imagination in their research and experiments?

a) Some scientists obtain better results in their research than others because of their creativity and imagination.
b) As long as scientists use the scientific method, they do not need to use their creativity and imagination.
c) Creativity and imagination are ambiguous concepts and they have no place in science.

19. Atoms are building blocks of all matter, but it is not possible to see atoms' internal structure. So which of the following statements do you agree about scientists' knowledge about atoms?

a) As atoms' pictures can be drawn and their models are made, they know the exact structure of atoms.
b) Even if atoms are very small, scientists discover their real structure with the experiments they conducted.
c) Even if atoms cannot be seen, thanks to experiments and observations, information about their structure can be obtained.

20. After living on earth for a long time, dinosaurs disappeared 65 million years ago. Which of the following statements about how much scientists are sure of their real appearance do you agree?

a) Scientists can be sure about the appearance of well-known dinosaurs like T-Rex and dinosaurs whose bones are found in abundance.
b) Based on bone and fossil findings and also with some imagination, they can only make comments about how dinosaurs looked.
c) Thanks to advancements in technology, the real appearance of dinosaurs will certainly be determined in the future if not today.

21. There are many fault lines that pass through Turkey. As a result of studies conducted about earthquakes, scientists think that in the near future there may be an earthquake in the Marmara Sea. However, they disagree on the time and severity of a possible earthquake. Why do you think scientists have different opinions even though they have the same information?

a) Scientists have different creativity and imagination and because of this, they always have differences in their opinions.
b) Earthquake research is relatively new and because of this, they have different opinions.
c) They have different opinions, because there aren't enough seismographs (tools that measure severity of an earthquake).

22. Whether cell phones cause cancer or not is being debated. Some researchers argue that extensive use of cell phones may cause cancer, while others could not find a relationship between cancer and cell phones. What do you think is the reason for this conflicting situation?

a) If scientists compare and discuss the data they collected, they will always arrive at the same conclusion and the conflicts will disappear.
b) It is normal for scientists to come to different conclusions about a subject. New studies may support one of these conclusions more so than others.
c) If scientists apply the scientific method correctly, they will always arrive at the same conclusions and these types of conflicts will not happen.

23. There are models such as cell model, DNA model, and atom model that are being used in science courses. Scientists use and produce various models as they investigate the nature. How much do you think these models reflect the reality?

a) The models being used in schools may be simple, but the models that scientists make exactly reflect the reality.
b) If enough attention is given to details, models will perfectly align with the reality.
c) Models are limited with the assumptions, creativity and means of people who created them and they never exactly reflect the reality.

24. In which of the stage/stages of research such as planning, experimenting, observing, analyzing data, interpreting data, reporting results do you think scientists use their creativity and imagination?

a) Scientists use their creativity and imagination more or less in every stage of their research.
b) Creativity and imagination is used in technological work and development of new products rather than science.
c) Scientific method is evident and there is no need for creativity and imagination in its application.

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 476–486
https://dx.doi.org/10.21449/ijate.616617

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

Practices, Challenges and Perceived Influence of Classroom Assessment on Mathematics Instruction

Isaac Buabeng [1,*], Amoah Barnabas Atingane [2], Isaac Amoako [3]

[1,2] Department of Basic Education, University of Cape Coast, Ghana
[3] Department of Education and Psychology, University of Cape Coast, Ghana

ARTICLE HISTORY

Received: 20 March 2019
Revised: 07 August 2019
Accepted: 22 August 2019

KEYWORDS

Classroom Assessment,
Instruction,
Feedback,
Assessment Practices,
Mathematics Teachers,
Junior High School

Abstract: Assessment is a powerful tool for raising the standards of teaching and learning of mathematics at the junior high school level. This study therefore explored the perceived influence of assessment on the teaching and learning of mathematics in junior high schools of OLA Circuit in Cape Coast Metropolitan area. The research design used for the study is a concurrent triangulation mixed method design. A simple random sampling technique was used to select four (4) public junior high schools out of eight (8) schools in the circuit. A multi-stage sampling procedure was employed to select the schools and participants for the study. A total of 134 participants comprising 15 teachers and 119 students participated in the study. The data for the study were mainly collected through questionnaires and interviews. Findings of the study revealed that class exercise, homework, and trial work were the most common mode of assessment used by teachers during mathematics instruction. Again, the study discovered that teachers faced some challenges in the implementation of classroom assessment. The study therefore makes certain recommendations likely to improve on the quality of assessment practices in mathematics classrooms in the focal schools.

1. INTRODUCTION

The essential purpose of education is to help the individual to be able to use their learning and their own mind as the anvil for creating new ideas, processes, gadgets and appliances (Curriculum Research and Development Division [CRDD], 2011). Mathematics is one of the essential areas of learning. According to the CRDD (2012),

> ‘‘today’s world demands that young people should be able to use numbers competently, read and interpret numeral data, reason logically, as well as communicate effectively with other people using accurate mathematical data and interpretations” (p. 3).

It is due to this that mathematics has been considered as one of the core subjects in the basic and the second cycle school curriculums in Ghana. However, the teaching and learning of mathematics at the Junior High School (JHS) level cannot be meaningful if students are taught only to repeat what is taught in school without giving them the opportunity to engage in critical

CONTACT: Isaac Buabeng ✉ ibuabeng@ucc.edu.gh Department of Basic Education, University of Cape Coast, Ghana

ISSN-e: 2148-7456 /© IJATE 2019

productive thinking and application of their knowledge to variety of situations while they are still in school.

For example, available statistics from the Cape Coast Metropolitan Education Directorate show that students' performance in mathematics is relatively low as compared to the other core subjects. For example, in 2010, the number of registered candidates who obtained passes (grades 1- 6) in mathematics in the Basic Education Certificate Examination (BECE) in the Metropolis was 36.7% as compared to 60.3%, 47.9% and 52.4% for English Language, Integrated Science and Social Studies respectively (Cape Coast Metropolitan Education Directorate, 2010). Similarly, in 2012, there was 40% number of candidates who obtained passes (grades 1 – 6) in Mathematics in the BECE as against 62%, 41% and 49% for English Language, Integrated Science and Social Studies respectively (Cape Coast Metropolitan Education Directorate, 2012). Again, in 2014, 60.95% number of candidates obtained passes (grades 1 – 6) in Mathematics in the BECE as compared to 76.36%, 70.49% and 62.88% for English Language, Integrated Science and Social Studies respectively (Cape Coast Metropolitan Education Directorate, 2014).

A careful look at the statistics above indicates that, students' performance of mathematics has been increasing over the years, however, the rate of increment is not substantial in comparison with the other core subjects such as integrated science, social studies and English. It has been argued that formative assessment practices serve the purpose of improving classroom instruction with subsequent effect on enhancing performance (Amoako, 2018). Also, it is expected that school-based assessment with particular emphasis on formative assessment will help teachers and pupils to achieve the objectives of the syllabus and consequently raise the standard of mathematics learning in the country (CRDD, 2011). Considering that the CRDD requires that all teachers incorporate formative assessment into their teaching due to its perceived benefits, it is curious as to why the mathematics performance of JHS students in mathematics within the Cape Coast Metropolis is not experiencing great gains. Could it be that teachers are not engaging in formative assessment practices? Or could it be due to ineffective assessment practices? In view of these and many other nagging questions, the authors investigated the kind of assessment modes (tools) and format that teachers use to drive instruction of mathematics in the area as well as any possible challenges that they face in the implementation of the various assessment procedures such as captured in the JHS mathematics syllabus. The study was guided by the following research questions:

1. What mode of assessment do JHS teachers use during mathematics instruction?
2. What format of assessment do JHS mathematics teachers frequently use?
3. What is the assessment feedback practices of mathematics teachers in OLA circuit?
4. What is the perceived influence of assessment practices on mathematics instruction in the JHS?
5. What challenges do mathematics teachers face during the implementation of assessment procedure in the classroom?

2. METHOD

Concurrent triangulation mixed method design was used for the study. This research strategy can be identified by its use of one data collection phase, during which both quantitative and qualitative data are collected simultaneously for the purpose of verification of information received (Creswell, 2003). For the purpose of this study, all JHSs in the Cape Coast metropolis were targeted. However, for efficiency of investigation, OLA circuit having eight (8) JHSs became the accessible population. The number of teachers within the circuit was estimated to be 70, made up of 29 males and 41 females whereas the number of students was also estimated to be 928, made up of 398 boys and 530 girls.

In selecting the samples, a multi-staged sampling procedure was used. At the first stage, purposive sampling procedure was used to select OLA circuit. The circuit was selected because it has most of the JHS within the Metropolis. On the second stage, random sampling method was used to select four JHSs out of a total of eight JHSs within the circuit. Using Krejcie and Morgan sampling size determination specifications (Sarantakos, 2005), a total of 119 students which was made up of 57(47.9%) males and 62(52.1%) females from the four randomly selected public JHSs participated in the study. Convenient sampling was also utilized to engage all the mathematics teachers from the selected schools, who were at post during the time of the study. This procedure was used to allow teachers who were ready and willing to participate in the conduct of the study to be selected. In all 15 mathematics teachers participated in the study.

Data for the study were obtained from two main sources, questionnaires for students and teachers, and interviews with teachers. The questionnaires were administered to both teachers and students while the interview was administered to the teachers. Only teachers were interviewed and not students because teachers make use of assessment procedures and hence would be able to tell how it affect instructions. The questionnaire (with overall Cronbach Alpha estimate of .72) was a four-point Likert scale with extreme responses of “Strongly Agree to Strongly Disagree.” The interview guide was semi-structured in nature which allowed the researchers to explore other issues as emerged from participant’s responses. The quantitative data were analyzed using mean and standard deviation whereas thematic approach was adopted for the qualitative data. Where quotes are used within the body of the results, they were chosen because they were representative of the statements by most of the respondents.

3. RESULTS

The results are presented as guided by the research questions which underpinned this study. In the next sections, we present the results of the research questions.

3.1. Mode of Assessment used by Teachers

Research question one sought to find out the mode of assessment that JHS teachers use in the classroom during mathematics instructions. Summary of the analysis is presented in Table 1.

Table 1. Teachers’ Views on the Mode of Assessment that they Commonly Use (N = 15)

Assessment Tool	Mean	SD
Class test	2.27	0.46
Class exercise	3.67	0.49
Homework	3.73	0.46
Group work	2.10	0.26
Project work	1.40	0.51
Trial work during lessons	3.53	0.64
Average scores	2.78	0.47

Mean Range: Not used (0.4–1.4), used occasionally (1.5–2.4), used often (2.5–3.4); used very often (3.5 – 4.4)

Table 1 shows that teachers often use variety of assessment modes in the classroom to assess students’ progress in mathematics. This is evident by the average mean score (*M*= 2.78, *SD* = .47). As shown in Table 1, class exercises (M=3.67, SD=0.56), homework (M=3.73, SD=0.46), and trial work (M=3.53, SD=0.64) were the modes of assessment often used by teachers.

Most of the teachers interviewed confirmed the above results when they asserted that they use more of the class test, class exercise, homework and trial work since these tools are prescribed in the school-based assessment guide. However, most of the teachers admitted in the interview that they do not use projects and group work as expected. Two teachers commented as follows:

> *I do not use project because it is not easy to find project topics for mathematics as compared to subjects like Integrated Science. [Teacher 'A']*
> *The students don't like working in groups and when you give them group work, they rather make noise instead of doing the work. [Teacher 'C']*

The second research question elicited from the students, their views about the assessment modes that are commonly used by their mathematics teachers. The results of Table 2 shows that the students held similar views as the teachers with regards to how often the named assessment modes were used in mathematics in their schools. The average mean score and standard deviation were 2.85 and 0.79 respectively. This indicates that the students view the assessment modes as 'used often' by their teachers.

Table 2. Students' Views on Assessment Modes that are Commonly used by their Teachers (N=119)

Assessment Tool	Mean	Std. dev.
Class test	2.50	0.74
Class exercise	3.73	0.56
Homework	3.47	0.74
Group work	2.40	0.92
Project work	1.81	0.98
Trial work during lessons	3.48	0.72
Average scores	2.85	0.79

Mean Range: Not used (0.4–1.4), used occasionally (1.5–2.4), used often (2.5–3.4); used very often (3.5 – 4.4).

An examination of the individual items points to the same results as seen in Table 3. Therefore, the students share similar views as their teachers.

3.2. Format of Assessment

Research question two sought to investigate the commonly used assessment format by mathematics teachers in the circuit. Summary of the analysis is shown in Table 3.

Table 3. Assessment Formats Commonly used by Mathematics Teachers (N = 15)

Assessment Format	Mean	Std. dev.
Essay type	3.47	0.74
Multiple choice	2.27	0.59
True/false	1.53	0.74
Matching items	1.67	0.62
Completion items	1.80	0.77
Average scores	2.15	0.69

Mean Range: not used (0.4–1.4), used occasionally (1.5–2.4), used often (2.5–3.4); used very often (3.5 – 4.4).

The average mean scores (M = 2.15, SD = .69) as shown in Table 3 indicate respondents' agreement that the listed assessment format are used by mathematics teachers in the OLA circuit occasionally. Table 3, further indicates that essay type questions were used very often (M=3.47, SD=0.74) while the rest of the formats, thus multiple choice (M=2.27, SD=0.59), true/false (M=1.53, SD=0.74), matching items (M=1.67, SD=0.62) and completion items (M=1.80, SD=0.77) were all used occasionally'. During the interview, most of the teachers acknowledged that both essay and multiple-choice type questions are prescribed for use at the junior high school level, but they (teachers) like using essay type questions since it easy to craft essay questions as compared to multiple choice questions.

Table 4, present summary of the analysis on student's opinion about assessment format commonly used by mathematics teachers.

Table 4. Students' Views on their Teachers use of Assessment Formats (N= 119)

Assessment Format	Mean	Std. dev.
Essay type	3.42	0.73
Multiple choice	2.37	0.78
True/false	1.66	0.92
Matching items	2.04	0.85
Completion items	2.20	0.87
Average scores	2.34	0.83

Mean Range: Not used (0.4–1.4), used occasionally (1.5–2.4), used often (2.5–3.4); used very often (3.5 – 4.4).

From Table 4, the average mean score and standard deviation were 2.34 and 0.83 respectively. This means that the students generally viewed the named assessment formats as 'used occasionally'. Generally, this is a confirmation of the views expressed by the mathematics teachers in Table 3. Also, a critical study of the individual items reveals similar trends as the views of the teachers in Table 5 and in the interview.

3.3. Assessment Feedback Practices of Mathematics Teachers

Research question three sought to solicit responses from both teachers and students about the promptness of teachers when it comes to providing assessment feedback and how they do it. Summary of the analysis is shown in Table 5.

Table 5. Teachers' Views on Assessment Feedback Practices (N=15)

Feedback Practice	Mean	Std.dev
I mark students work and quickly gives it back to them	3.13	0.52
I revise assessment task with my students	3.20	0.68
I rank my students test results	2.23	0.83
I motivate students who perform well in Mathematics	2.60	0.82
I provide written comments along with students' marks	3.20	0.68
I point out my students' weaknesses to them	3.13	0.64
I talk to students about how they can improve their Performance	2.73	0.80
I organize remedial teaching for those who get low marks	1.67	0.72
I use assessment results to provide guidance to my students	2.53	0.64
Average scores	2.72	0.67

Mean Range: Never (0.4 – 1.4), sometimes (1.5 – 2.4), most of the times (2.5 – 3.4); always (3.5 –4.4).

The average mean score and standard deviation in Table 5 were 2.72 and 0.67 respectively. This generally means that the teachers 'most of the time' carry out the stated feedback practices in their schools. For instance, the individual item analysis on Table 5 further shows that the teachers most of the time mark their students work (M=3.13, SD=0.52) and revise assessment task with them (M=3.20, SD=0.68). Similarly, the results show that the teachers most of the time provide written comments along with students' marks (M=3.20, SD=0.68), and point out students' weaknesses to them (M=3.13, SD=0.64).

During the interview, most of the teachers asserted that they regularly mark and revise their students work with them. This shows the efforts made by the teacher in using assessment to help students know their learning progress. The teachers interviewed admitted that they do not rank students test results except the end of term exams.

Table 6 provides summary of the analysis about students' views on assessment feedback practices of mathematics teachers.

Table 6. Students' Opinion on Assessment Feedback Practices of Mathematics Teachers (N=119)

Feedback Practice	Mean	Std.dev
Teacher marks our work and gives it back quickly	2.96	0.75
Teacher revises assessment task with us	2.63	0.86
Teacher ranks our test results	2.34	0.47
Teacher motivates students who perform well in mathematics	2.53	0.64
Teacher provides written comments along our marks	2.92	0.77
Teacher points our weaknesses to us	2.61	0.70
Teacher talks to us about how we can improve our Performance	2.62	0.75
Teacher organizes remedial teaching for those who get low marks	1.72	0.97
Teacher uses assessment results to provide guidance to us	2.53	0.64
Average scores	2.55	0.84

Mean Range: Never (0.4 – 1.4), sometimes (1.5 – 2.4), most of the times (2.5 – 3.4); always (3.5 – 4.4).

The average mean score and standard deviation of Table 6 were 2.55 and 0.84. This means that the stated feedback practices 'most of the time' were carried out in the schools. This clearly confirms the views expressed by the teachers about the occurrence of the stated assessment feedback practices as contained in Table 5 and in the interview.

3.4. Impacts of Assessment on Mathematics Instruction

Research question four was intended to find out the perceived impact of assessment practices on mathematics instruction. Summary of the teachers' responses is presented in Table 7.

Table 7. Teachers' Views on the Impacts of Assessment on Mathematics Instruction (N=15)

Impacts of Assessment	Mean	Std. dev
It helps me to identify and improve the weaknesses of my students	3.47	0.64
It develops my students' confidence in mathematics	3.07	0.46
It develops my students' interest in mathematics	3.20	0.56
It helps me to monitor my students learning progress	3.60	0.51
It helps me to involve my students in my lessons	3.53	0.52
It helps me to identify students who need special attention in learning mathematics	3.32	0.59
It helps me to know if my lesson objectives are being achieved	3.53	0.64
It helps me in putting my students into appropriate learning groups	3.27	0.59
Average scores	3.37	0.56

Mean Range: Strongly disagree (0.4 – 1.4), disagree (1.5 – 2.4), agree (2.5 - 3.4); strongly agree (3.5 – 4.4).

The average mean score and standard deviation of Table 7, being 3.37 and 0.56 respectively indicate that the teachers generally 'agreed' that assessment has an impact on mathematics instruction. A critical study of the individual statements show that the teachers strongly agreed that assessment helps them to identify and improve the weaknesses of their students (M=3.47, SD=0.64). Also, the teachers agreed that assessment develops students' confidence in mathematics (M=3.07, SD=0.46), and it develops students' interest in mathematics (M=3.20, SD=0.56). Again, the results revealed that the teachers strongly agreed that assessment helps them to monitor their students learning progress (M= 3.60, SD=0.51), and to involve their students in their lessons (M=3.53, SD=0.52). Furthermore, Table 7 shows that the teachers agreed that with assessment, they were able to identify students who need special attention (M=3.32, SD=0.59). The teachers also strongly agreed (M=3.53, SD=0.64) that through assessment they were able to know if their lesson objectives were achieved. In the interview, many of the teachers acknowledged that assessment actually impact teaching and learning of mathematics. Two teachers commented:

> *Assessment helps me to know if my lesson was well taught and to know the next thing to do. [Teacher 'D']*
>
> *With assessment, I am able to collect information that enable me make decisions about my students learning progress and my own teaching strategies. [Teacher 'C']*

Table 8 shows the summary of analysis concerning students views about impact of assessment on students learning.

Table 8. Students' Views on the impact of assessment (N=119)

Impact of Assessment	Mean	Std. dev
It helps me to identify and improve my weaknesses	3.53	0.70
It develops my confidence in mathematics	2.82	1.03
It develops my interest in learning mathematics	2.87	0.99
It helps me to monitor my learning progress	3.23	0.73
It helps me to know what to learn	3.54	0.74
Average scores	3.20	0.84

Mean Range: Strongly disagree (0.4 – 1.4), disagree (1.5 – 2.4), agree (2.5 - 3.4); strongly agree (3.5 – 4.4).

The average mean score (M= 3.20, SD = .84) as shown by Table 8 implies that the students largely agree with the teachers on the statements about the role assessment plays on teaching and learning of mathematics in their schools. For instance, Table 8 shows that the students strongly agree that through assessment they were able to identify and improve their weaknesses in mathematics (M=3.53, SD=0.70). The students' views largely confirm the views of the teachers as presented in Table 7 and in the interview.

3.5. Challenges Associated with Implementation of Assessment Procedures

The last research question sought to find out the challenges that the teachers perceive to be hindering the quality of assessment of students in mathematics in their schools. Details of the challenges are shown in Table 9.

Table 9. Teachers' Perceived Challenges that constrain Quality Assessment Practices in Mathematics (N=15)

Challenge	Mean	Std. dev
The school has inadequate assessment materials	3.07	0.46
Assessment increase my workload	3.13	0.64
Assessment takes much of my time	3.13	0.64
I do not have adequate skills on assessment in mathematics	3.20	0.56
Some of my students do not submit their work for marking	3.07	0.59
My students' attendance to school is poor	3.13	0.74
Average scores	3.12	0.61

Mean Range: Strongly disagree (0.4 – 1.4), disagree (1.5 – 2.4), agree (2.5 - 3.4); strongly agree (3.5 – 4.4)

As shown in Table 9, the average mean score and standard deviation were 3.12 and 0.61 respectively. This indicates that the teachers largely 'agree' to the statements about the challenges that constrain quality assessment in mathematics in their schools. A study of the individual statements revealed that the teachers agree that their schools had inadequate assessment materials (M=3.07, SD=0.46), and assessment increases their workload (M=3.13, SD=0.64). Similarly, the results of Table 9 shows that the teachers agree with the assertions that they (the teachers) do not have adequate skills for assessing students in mathematics (M=3.20, SD=0.56), and some students do not submit their work for marking (M=3.07, SD=0.59).

The information gathered through the interview conducted largely confirmed the results of the questionnaire. For instance, in the interview, most of the teachers reported inadequate materials for assessment, increase workload as well as failure of some students to submit their assessment task for marking due to poor attendance to school as some of the challenges to quality assessment practices in their schools. Here, the non-availability of materials such as SBA books, report cards, graph sheets and answer booklets greatly affected the assessment practice that were being carried out in these schools.

On the issue of increased workload, many of the teachers also complained that the mathematics syllabus is loaded besides they teach other subjects in addition to the mathematics. These suggest that the high workload makes them to pay less attention to assessment of their students learning. One of the teachers commented;

> *The teaching alone takes all my time and I will not be able to finish the syllabus if I am to engage in effective assessment practices like organizing remedial lessons for students who normally get low marks. [Teacher 'A']*

4. DISCUSSION

The study revealed that teachers often use variety of assessment modes in the classroom to assess students' progress in mathematics. The modes include class exercise, homework and trial work. However, when it comes the use of group work and project work, the teachers indicated that they do not use it at all. It is more likely that the teachers sideline of project work and group work might be as a result of lack of proper understanding of the usefulness of these methods or probably insufficient instructional time at their disposal. This practice actually deviates from CRDD (2011) directive that the performance of students in mathematics can best be assessed if the assessment is made on different test modes including projects, mental exercises, group exercises (cooperative learning exercises) and other practical activities. The present study discovery of the use of varied assessment procedures is in line with the findings of Kipkorir (2015) who discovered that mathematics teachers in the Nandi Central Sub-County, Kenya, have used multiple methods of assessment such as discourse, observation, students' self-assessment and peer assessment which have had massive turns on students learning of concepts in mathematics. Equally, other studies have also shown enormous significance of 'varied assessment modes in students learning (Birgin, 2011; Buhagiar, 2007).

Most teachers and the students agreed that essay type questions were the predominantly used assessment format. In an interview section with some of the teachers, they explained that, they often times use essay type of test because, it is easy to construct. It is known in literature that ideally, the purpose of the test, the difficulty level that the teacher anticipates and the characteristics of the test takers inform the appropriate format to use. In a situation whereby, teachers resort to just a particular test format because they perceive it to be easier when constructing such test, then it is more likely that the teachers lack adequate competencies in test construction. This result is in line with Quansah, Amoako and Ankomah (2018) who discovered that teachers in the Cape Coast Metropolis have limited skills in the construction of test items. Moreover, it could be possibly due to the fact that teachers have poor attitude when it comes to test construction and hence overreliance to a particular test format (Quansah & Amoako, 2018).

The study showed that teachers 'most of the time' carry out the stated feedback practices in their schools. This was actually corroborated by the responses of students on the questionnaire. However, there were few areas that teachers indicated that they normally do not get time to organize remedial classes for students as part of the feedback exercise. This excuse from the teachers might have been born out of the fact that they see feedback exercise as distinct from the teaching and learning encounter. This confirms the assertion made by Taras (2003) that the challenges with feedback are that teachers and students see feedback in isolation from other aspects of the teaching and learning process, and considers feedback to be primarily a teacher-owned endeavour.

It was evident from the results of the study that teachers and students alike perceive assessment as having impact on mathematics instruction. On the part of the teachers, they believe that assessment helps them to know of the lesson was well taught. It also helps them to know the weakness of the students. Students also use the assessment results to gauge their understanding of concepts taught by the teachers in class. This study finding as shown corroborates several other study findings in the literature. For example, Black and William (2010) that teachers can interpret and use assessment results to gauge whether the teaching has been successful in achieving its objective(s). Black and William added that the teacher may then use assessment results as the basis for giving advice on students learning or reviewing teaching. Again, Koloi-Keaikitse (2012) found that in order for teachers to diagnose students' needs, design and implement instructional interventions, evaluate students work, and assign grades, they (teachers) need continuous access to evidence of students learning arising from high-quality classroom assessment practices. In the context of classroom instruction, formative assessment practices help students to know whether they have understood the concept taught by the teacher or not, this serve a motivational role for extra effort on the part of the student (Amoako, 2018).

Finally, the study discovered some challenges that frustrate classroom assessment practices to include; inadequate assessment materials, high workload of teachers and poor attendance of students to school among other minor ones. The teachers' views are in line with the opinion of Tamakloe, Amedahe, and Attah (1996) that assessment especially continuous assessment is time consuming as teachers have to construct their assessment tasks, administer them, grade the scores, have the scores recorded and then carry out revision with the students. Tamakloe, Amedahe and Attah added that assessment increases the workload of teachers.

5. CONCLUSION

Based on the findings of the study it can be concluded that mathematics teachers in junior high schools of OLA Circuit tend to use more of class test, homework and trial work in assessing students learning to the neglect of group work and project work. This situation is more likely to deny students the benefits of knowledge sharing (learning from peers), in this case teaching and learning of mathematics would be done in abstract.

Again, it can be concluded that there is undue emphasis on the use of essay type questions in the schools which make it difficult to adequately prepare the students for the BECE in which essay questions and objectives/multiple choice questions are weighed equally. Moreover, assessment practices have an impact on classroom mathematics instruction which ranges from promoting involvement of students in mathematics lessons to increasing teachers' pedagogical effectiveness. Assessment is therefore a powerful tool for enhancing effective teaching and learning of mathematics.

Finally, despite the generally acclaimed benefits of assessment, certain challenges such as inadequate assessment materials, high workload of teachers and poor attendance of students to school tend to frustrate the positive impact of assessment on the teaching and learning of mathematics. Based on the findings from the study, it is recommended that Head teachers should ensure that as part of the school base assessment (SBA) procedures, teachers' pay particular attention to project work and group work. These procedures have the tendency to encourage peer tutoring among students which augment classroom instruction. In addition, head teachers could collaborate with the GES to organize regular in-service programmes for the mathematics teachers to constantly update their knowledge, skills and attitudes toward assessment. Head teachers and the teachers also need to liaise with the educational authorities and philanthropist to provide material necessary for assessment practices in the various schools.

ORCID

Isaac Buabeng https://orcid.org/0000-0003-4569-087X

6. REFERENCES

Amoako, I. (2018). Formative assessment practices among distance education tutors in Ghana. *African Journal of Teacher Education, 7(*3), 22-36.

Asamoah-Gyimah, K. & Duodu, F. (2007). *Introduction to research methods in education.* Winneba: The Institute of Educational Development and Extension, University of Education, Winneba.

Birgin, O. (2011). Pre-service mathematics teachers' views on the use of portfolios in their education as an alternative assessment method. Educational Research and Reviews, *6*(11), 710-721.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2010). *Assessment for learning, putting it into practice.* London: Open University Press.

Buhagiar, M. A. (2007) Classroom assessment within the alternative assessment paradigm: revisiting the territory. *Curriculum Journal*, *18*(1), 39 – 56.

Cape Coast Metropolitan Education Directorate (2010). Analysis of 2010 basic education certificate examination results. Cape Coast.

Cape Coast Metropolitan Education Directorate (2012). Analysis of 2012 basic education certificate examination results. Cape Coast.

Cape Coast Metropolitan Education Directorate (2014). Analysis of 2014 basic education certificate examination results. Cape Coast.

Curriculum Research and Development Division (2012). *National syllabus for mathematics: Junior high school 1- 3.* Accra: Ministry of Education.

Curriculum Research and Development Division (2011). *Teachers hand book on school-based assessment for junior high schools (Mathematic).* Accra: Ministry of Education.

Dillard, J. (2013). *Five most important methods for statistical data analysis.* Retrieved, from http://www.bigskyassociates.com/blog/bid/356764/5-Most-Methods-ForStatistical-DataAnalisis

Holt, L.C., & Kysilka, M. (2006). *Instructional patterns: Strategies for maximizing students learning.* California: Sage Publication, Inc.

Kipkorir, K. E. (2015). Classroom assessment practices by mathematics teachers in secondary schools of Kenya. Unpublished Masters thesis, University of Nairobi, Kenya.

Larry, H. C., & Kysilka, M. (2006). *Instructional patterns: Strategies for maximizing students' learning.* California: Sage Publication, Inc.

Ornstein, A. C., & Lasley, T. J. (2000). *Strategies for effective teaching.* United States of America: The McGraw Company Inc.

Quansah, F. (2017). The use of Cronbach alpha reliability estimate in research among students in public universities in Ghana. *African Journal of Teacher Education, 6*(1), 56-64.

Quansah, F., Amoako, I., & Ankomah, F. (2018). Teachers' Test Construction Skills in Senior High Schools in Ghana: Document Analysis. *International Journal of Assessment Tools in Education*, *6*(1), 1-8.

Salaria, N. (2012). *Meaning of the term descriptive survey research method. International Journal of Transformations in Business Management,1*(6)*,* Apr-Jun. Retrieved from http://ijtbm.com/images/short_pdf/Apr_2012_NEERU%2520SALARIA%25202.pdf.

West African Examination Council (2013). *Basic education certificate examination: Chief examiners' report.* Accra: West African Examination council.

Quansah, F., Amoako, I. (2018). Attitude of Senior High School (SHS) teachers towards test construction: Developing and validating a standardized instrument. *Research on Humanities and Social Sciences, 8*(1), 25-30.

Sarantakos, S. (2005*). Social research* (3rd ed.). New York: Palgrave Macmillan

Tamakloe, E. K., Amedahe, F. K., & Attah, E. T. (1996). *Principles and methods of teaching.* Accra: Black Mask Ltd.

Taras, M. (2003). To feedback or not to feedback in student self-assessment. *Assessment and Evaluation in Higher Education, 28*(5), 549- 565.

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 487–505
https://dx.doi.org/10.21449/ijate.616795

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

Examination of Student Growth Using Gain Score and Categorical Growth Models

Hatice Cigdem Yavuz [1,*], **Ömer Kutlu** [2]

[1]Cukurova University, Faculty of Education, 01330, Sarıçam/Adana, Turkey
[2]Ankara University, Faculty of Educational Sciences, 06590 Çankaya/Ankara, Turkey

ARTICLE HISTORY

Received: 30 May 2019

Revised: 08 August 2019

Accepted: 31 August 2019

KEYWORDS

Gain score model,
Categorical growth model,
Student growth,
Multilevel modeling

Abstract: In this study, gain score, and categorical growth models were used to examine the role of student (gender and socioeconomic level) and school characteristics (school size and school resources) in the student growth on comprehension skills in language. The participants of this study were 2,416 sixth-grade students in 2011 who became seventh-grade students in 2012. The data was collected through two achievement tests, student and school questionnaires. Two achievement tests were calibrated using the Rasch Model and were scaled using the concurrent estimation method. Moreover, the cut-off scores of these tests were determined by using the bookmark method. Students' growth was modelled with the gain score and categorical growth models. All data was analyzed using multilevel models. Results showed that some students did not achieve sufficient gains to advance to higher performance levels. Although some schools' average gains were higher, their performance was still not significant enough in terms of tests' standards. Moreover, the analyses demonstrated that the student gain scores and growth categories varied significantly among the schools. In addition, the study was able to determine student and school characteristics that have an impact on the students' gain scores and categorical growth. Given the different aspects gained about students' performance with these models, it is recommended to utilize different growth models in schools.

1. INTRODUCTION

The widespread use of assessments in education which focus on students' performances determined from a single time point is a point of contention in the field of assessment studies (Betebenner & Linn, 2009). The reason for this is because the information obtained from such assessments is limited. Within the scope of this limited information in question, multiple questions arise concerning the validity of classification of students, determining of their performance levels and considering students with learning difficulties, as well as inference from teachers and schools (Laird, 2008). Furthermore, together with these assessments, education shareholders are able to see the growth of students, and they may be able to discern whether the growth in question is in accordance with the standards as well (Yen, 2007; cited in Betebenner & Linn, 2009). In this sense, assessments that measure development can provide more clear

CONTACT: Hatice Cigdem YAVUZ ✉ hcyavuz@cu.edu.tr Cukurova University, Faculty of Education, 01330, Sarıçam/Adana, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

information regarding school effectiveness and student achievement (Heck, 2006). Indeed, applications that measure growth have been effectively used in many educational systems (e.g. Assessment Agency, 2008; NCLB [No Child Left Behind], 2002; U.S. Department of Education, 2010) for many years and are ever increasing in their importance (Briggs & Betebenner, 2009).

Studies focusing on student growth in the literature predominately concern themselves with comprehension skills in language and reading comprehension skills (Herbers, Cutuli, Supkoff, Heistad, Chan, Hinz, & Masten, 2012; Hughes, Luo, Kwo, & Loyd, 2008; McCoach, O'Connell, Reis, & Levitt, 2006; Skibbe, Connor, Morrison, & Jewkes, 2010). The reason for this is that reading comprehension and language skills are one of the most important fields in terms of educational accountability (Shin, Davison, Long, Chan, & Heistad, 2013). Moreover, these skills can play a significant role in student academic achievement in other subjects (Arnold & Doctoroff, 2003; Crawford, Tindal, & Steiber, 2001). In this sense, academic achievement at higher grade levels of students who have difficulty in reading comprehension are also affected negatively by such deficiencies in upcoming grades (Crawford et al., 2001; Herbers et al., 2012). In this context, it can be stated that the linguistic skills are of great importance for students’ academic lives.

As in academic achievement, gender differences are presented in the academic growth of students as well. In the related literature, it is indicated by many studies that the growth of female students in language and reading comprehension skills generally surpasses that of male students (Denton & West, 2002; Husain & Millimet, 2009; Kurdek & Sinclair, 2001). In particular, studies pointed out that the students' initial level of reading comprehension is different according to gender (Morgan, Farkas, & Wu, 2011). Another crucial characteristic determining student success is the fact that the socioeconomic level of the student also plays an important role in the academic growth (McCoach et al., 2006; Nese, Biancarosa, Anderson, Lai, Alonzo, & Tindal, 2012). According to Shin et al. (2013), students coming from lower socioeconomic backgrounds have shown lower academic achievement in the fields of comprehension skills in language and reading than other students. Similarly, it has been seen that the growth of the students coming from lower socioeconomic backgrounds is slower than the other students (Palardy, 2008).

Another area of interest is the role of school size and school resources which are among the school characteristics examined in student growth varies according to school types in different countries and different studies (Hanushek, 2006; Stevensen, 2006). According to studies in the literature, the effect of school size on student achievement may vary in degree or even its direction; however, it’s also possible that such an effect may not be observed (Stevenson, 2006). The same situation is also observed on student growth (Heck, 2006; Palardy, 2008). In the literature, the effect of school resources on the student's academic achievement has been studied effectively for many years (Hanushek, 2006). Studies in the literature reveal different results about the effects of school resources on student achievement (Krueger & Lindahl, 2001). In the literature, some studies reveal the positive effect of school resources on student growth (Cheti & Birgitta, 2012; Palardy, 2008), whereas others have not found a significant effect (Glewwe, Hanushek, Humpage, & Ravina, 2011).

Only a limited number of studies that pursue data on the academic growth of students have looked (Ergin-Aydemir & Sünbül, 2016; Bursal, 2013; Yapar, 2014) at Turkey. These studies did not focus on students' linguistic skills. Of these studies, only Yapar (2014) conducted a study on student growth in English reading skills. Although student growth is not monitored in reading comprehension skills in Turkey, there are studies that examine students' status in this field according to student and school characteristics (Erman-Aslanoğlu and Kutlu, 2015; Kutlu, Yıldırım, Bilican, & Kumandaş, 2011; Güzle-Kayır & Erdoğan, 2015; Özer-Özkan & Doğan, 2013). The results of these studies show that student and school characteristics effect students' reading comprehension. Large-scale assessments conducted in Turkey (e.g.; ABIDE [Monitoring and Evaluation of Academic Skills], PISA [The Programme for International Student Assessment] and PIRLS [Progress in International Reading Literacy Study]) show that Turkish students do not excel in reading comprehension (Organization for Economic Cooperation and Development [OECD], 2014a, 2016; Mullis, Martin, Gonzalez, & Kennedy, 2001). The findings of these studies seem to agree that student performance in reading comprehension might vary based on the student and school characteristics. However, the role of school and student characteristics in Turkish students' growth cannot be determined because there are no studies regarding monitoring students in Turkey. Nevertheless, outside of Turkey many studies in the literature have monitored the growth of comprehension skills in language and reading comprehension skills of students in the context of different student and school characteristics (McCoach et al., 2006; Palardy, 2008; Skibbe et al., 2010; Shin et al., 2013).

Since growth can be measured according to gains and/or norms, it should be asked whether growth is intended to be measured based on performance standards or groups (Gong, 2004). The change in student performance in the gain score model can be seen with a calculation made by subtracting the score obtained in the previous years from the score obtained in the relevant year (Welch, Dunbar, & Rickels, 2016). In the categorical growth model, student growth is converted into the performance levels corresponding to the student's scores and inferences are made based on these performance levels. In the context of this study, the modeling of growth in comprehension skills in language over a year has been modeled according to the gain scores and the performance levels. By including student and school characteristics into these two different student growth models within the framework of educational accountability, this study aims to determine the effects of these characteristics on student growth.

In Turkey, there is a lack of sufficient relevant data on students' academic growth and the growth of different student groups. Thus, how the role of students' performance levels changed in one year, and the role of school and student characteristics in this change is not known. Assessments measuring students' performance, which are used instead of student growth models, do not enable the Turkish education system to grow. School shareholders are excluded from the educational accountability system since there is no data/information source in the Turkish education system for comparison based on certain standards of accountability (Nayır, 2013). In this sense, some researchers have concluded the practices related to accountability in the Turkish education system are insufficient (Türkoğlu & Aypay, 2015) and these practices are not informative for the shareholders.

In this study, gain score, and categorical growth models were used to examine the role of student (gender and socioeconomic level) and school characteristics (school size and school resources) in the student growth on comprehension skills in the language. In other words, this study aims to determine students' gain scores and at the same time, to monitor growth according to performance levels. In this context, answers to the following research questions are sought in the study: (i) What are the frequencies of schools in the gain scores and in growth categories?; (ii) What are the effects of the student gender and socioeconomic status as well as school characteristics such as school size and school resources on student gain scores?; (iii) What are the effects of the student gender and socioeconomic level and school characteristics such as school size and school resources on student growth categories?

2. METHOD

2.1. Study Group

The participants of this study are composed of 52 schools and 2146 students (52.21% female) which participated in the "Learning Level Research (LLR)" LLR-1 and LLR-3 (Ayral, Özdemir, & Sadıç, 2011) that Altındağ Guidance and Research Center carried over the 2011-

2012 and the 2012-2013 academic years during the spring term. The students included in the study were in the sixth grade in the 2011-2012 academic year in Altındağ district of Ankara, and the same students who continued as seventh graders in the 2012-2013 academic year.

2.2. Data Collection Tools

2.2.1. *The achievement tests*

Two achievement tests were applied in LLR 1 and LLR 3 projects. The test used in LLR 1 was applied as 18 multiple choice items with four choices in the 2010-2011 academic year while the other one in LLR 3 were applied as 54 multiple choice items with four choices in the 2011-2012 academic year. Together with two field experts, a total of five experts were employed in the development of the test which was applied in LLR 1. Together with four field experts, a total of 10 experts were employed in the development of the test, which was implemented in the LLR 3. The items, which were designed to measure comprehension skills in language, were placed in the achievement tests. In the development of the test, PISA, PIRLS studies and primary school curriculums were taken into consideration in determining the skills to be dealt with in the tests (Ayral et al., 2011). Before creating the final forms of the tests, a pilot study was conducted, and the items were revised based on item statistics analysis (for more information; see Ayral et al., 2011).

In this study, two new sub-tests were created by selecting items from the tests in the LLR 1 and LLR 3. The item selection of both tests was conducted based on the Rasch Model of the Item Response Theory (IRT) after checking all necessary assumptions. While selecting the specific items, attention was paid to the contents of the items, the item and item-fit statistics. In this context, 13 items were selected each from LLR 1 and LLR 3 while three of which were pseudo-common items.

For the newly created tests, the names NLLR1 and NLLR3 were used, respectively. After finalizing item selection and preparing new test forms, the performances of the students who participated in the LLR 1 and LLR 3 were re-estimated according to the answers they gave to the relevant items in NLLR1 and NLLR3. In this sense, new scores of the students were estimated within the scope of newly created test forms.

The KR-20 reliability coefficient was found to be .62 for both tests. The fact that there are a limited number of items in the tests may prevent the reliability to be higher. As shown in Figure 1, the information appears to be much greater in the theta range between −1.5 to +1.5 for the NLLR 1 and NLLR 3.

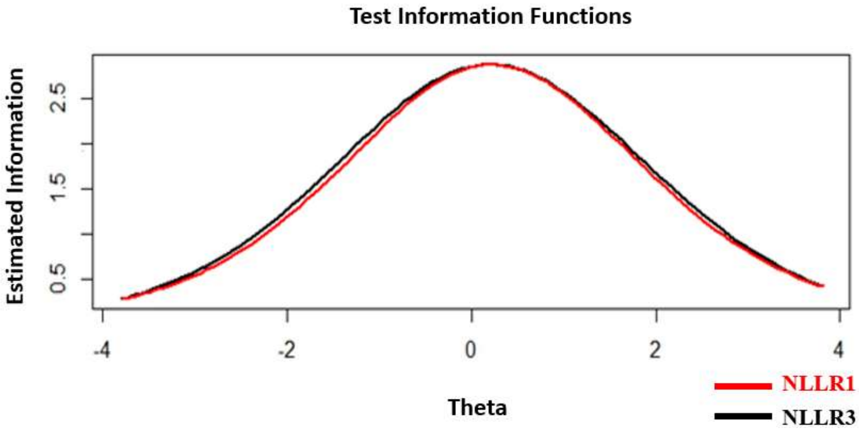


Figure 1. Test information functions for the NLLR 1 and NLLR 3

2.2.2. *Student and School Questionnaires*

In addition to the achievement tests, students were given questionnaires in LLR 1 and LLR 3. This study gathered information regarding the gender and socioeconomic level in the given questionnaire. For the socioeconomic level of the students, an index variable was created by using the education level of the parents of the students, the number of books at home and the per capita income in the family. Moreover, the size of the school and the school resources (physical resources) were taken into consideration as school characteristics. In the data set, the school size variable was accepted as the total number of students. Another index variable was created for school resources in which the number of classrooms, the number of laboratories, music rooms, painting rooms, and the number of gyms were identified. The index calculation is explained in detail at the upcoming section.

2.3. Procedure and Data Analysis

Before conducting the data analysis, the missing values in the data set were checked. Firstly, data belonging to students who did not have any data at both of the two measurement points were excluded from the data set. This data set was used in statistical procedures during the preparation of the data for analysis. In the analyses conducted in order to find answers to the research questions, the data set excluding students without questionnaire data was used. In addition, since the parameter values could not be produced without bias in groups with clusters consisting of fewer than 12 units (Browne & Draper, 2006), students in schools with fewer than 12 students were excluded from the data set as well. Within this context, analyses were carried out on a total of 2004 students for the research questions. With the final data set, the data imputation was not done due to full information maximum likelihood (FIML) estimation used for the analyses.

The following steps were followed in preparing the data set for analysis:

1. Item calibration: The items in the achievement tests were calibrated according to the Rasch model for this study after checking all necessary assumptions (unidimensionality, local independence and item fit). Item difficulties ranged from -.03 to 1.53 for NLLR 1, and .00 to 1.53 for NLLR 3, with the average difficulty being .00 for both tests, which means that the items were of moderate difficulty overall.

2. Selection of the items: First of all, three pseudo-common items were chosen from both tests. The selection of pseudo-common items is based on Luppescu (2005)'s selection criteria regarding items to be used in virtual equating. In this sense, cognitive levels, subject areas and difficulty levels of the items were taken into consideration in the selection of pseudo-common items. After the selection of pseudo-common items with priority, the items were selected in a way that the number of items in the subject fields of the new tests to be created would be equal, and have similar reliability coefficients and average difficulties. According to this, initially, five items with low item quality were removed from LLR 1. Thus, with the 10 left and three pseudo-common items, NLLR 1 was created. After that, a total of 10 items were selected from the LLR 3 that were in accordance with the items in the NLLR 1. Thus, NLLR 3 test was created. For NLLR 1 and 3, item calibration based on the Rasch model was done again.

3. Scaling of tests: The tests were scaled so that the scores obtained from the tests used in this study would be interpreted correctly. In this study, the scaling of the tests was conducted by concurrent estimation, which is one of the IRT methods. In the concurrent estimation, all parameters can be on the same scale since all parameters of the items in both tests are estimated concurrently in a single run (Hanson & Beguin, 2002). For this reason, these methods do not require any conversion between forms (Gonzalez & Wiberg, 2017).

4. Converting scores from the ability parameters into test scores: As two performance levels were determined for the tests created in this study, the obtained ability parameters of the students were converted into scores in the range of 0-200.

5. Standard setting for the tests: In this study, the cut-off scores for NLLR 1 and 3 were determined, as tests with determined standards should be used in order to employ the categorical

growth models. The bookmark method, which is among test-centered bookmark methods (Lewis, Mitzel, & Green, 1996), was used for this research.

According to the bookmark method, two performance levels namely "basic level" and "proficient level" were determined for NLLR 1 and 3. In this study, six female Turkish-language teachers working in a primary school located in the Mamak district of Ankara served as panelists during the standard-setting process. The average year of seniority of the panelists was 8.83, but it ranged between 4 and 18 years.

In the study, before the standard setting of the tests, panelists underwent training related to the standard setting, the bookmark method and the tasks expected from. Considering the characteristics of the tests used in the study and the information that should be given for the training of the panelists, an agenda was formed for the standard setting panel. The panel was conducted in parallel with this agenda which was determined for the standard setting. At the end of three rounds with the panelists regarding the setting of the standards, the panelists determined the cut-off scores for NLLR 1 as 101.23 and 98.78 for NLLR 3.

6. Calculation of gain scores: After the tests were scaled, the achievement scores of the students in the year 2011 were subtracted from the achievement scores of 2012.

7. Determining the growth categories: Categorical growth models are defined as student growth which is converted into categorical performance levels corresponding to the student's scores, and making inferences over performance levels. In this context, according to the determined cut-off scores of NLLR 1 and 3, students' test scores were converted into the performance levels in both tests. Afterwards, the change students displayed from NLLR 1 to NLLR 3 was categorized. For this purpose, the students who pass towards a higher performance level were coded as 3, students who remained at a high-performance level were coded as 2, students remained at low-performance level are coded as 1 and students whose performance level downgraded to a basic level were coded as 0. In this sense, it should be noted that this categorization was made based on dummy coding, not on an ordinal categorization.

8. Creation of indexes of socioeconomic level and school resources: In the creation of the index variable, the index formula (OECD, 2014b, p. 352) was employed. In the calculation of the values in the formula, principal component analysis was used for the school resources variables; mixed principal component analysis was used for the socioeconomic level variables as those mentioned variables were composed of continuous and categorical variables together. Mixed principal component analysis was conducted in the R program with the package named as "PCAmixdata" (Chavent, Kuentz, Labenne, Liquet, & Saracco, 2014).

Descriptive statistics were used in the first research question of the study and multilevel models were used separately for the second and third research questions. In the study, the first level was taken as the student level and the second level was taken as the school level. There are two reasons for employing these models in the study. The first one is that the data set used in the study is nested, and these models can analyze more than one level in these structures more reliably (Raudenbush & Bryk, 2002). The other reason is to take into account that other analyses, apart from multilevel models require the assumption that the observations are independent. However, the students are not randomly placed in schools and the multilevel models are able to eliminate said problem (Osborne, 2000).

In the second research question, multilevel models (one-way ANOVA, first-level random intercept and regression model in which means are outcomes) were used. Mplus 8 program was used in the analysis of these models. The Mplus 8 program is advantageous and strong in terms of parameter prediction (Muthén & Muthén, 1998; 2017). In the third research question, the hierarchical generalized linear model (HGLM) was used. The HGLMs are different from multilevel models because the dependent variable does not meet the normality assumption and they are more compatible in terms of distributions (Raudenbush, Bryk, Cheong, & Congdon, 2011). The stated analyses were carried out with HLM 7.03 program (Raudenbush et al., 2011). The Pratt index was calculated in order to determine the importance of predictor variables in predicting the dependent variable which is taken with multilevel models. Pratt index is used to calculate the relative importance of predictor variables (Liu, Zumbo, & Wu, 2014).

3. FINDINGS

3.1. Findings of the models using gain score model

Descriptive statistics related to the values calculated according to the gain score model of schools are given in Appendix 1 and the results based on statistics are given in Figure 2.

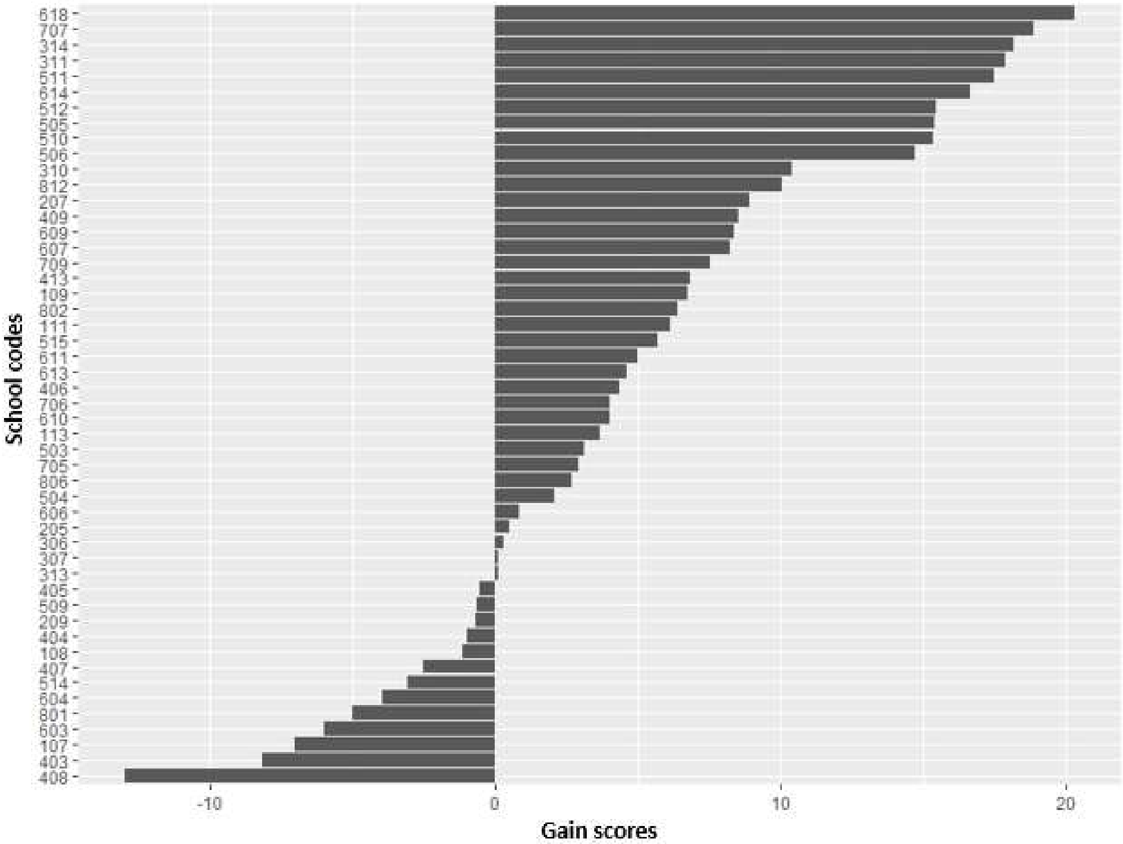


Figure 2. Frequency distribution of the average gain scores of schools

According to Figure 2, in terms of the gain model of the students who participated in the study, the school with the code of 618 has the highest average score, while the 408 coded school has the lowest average score. In addition, according to Appendix 1, the most homogenous school is the school with the code of 107, while the most heterogeneous school is the school with the code 707 in terms of gain scores.

3.2. Findings related to differences between schools in terms of gain scores

The results regarding the models established to examine the differences between schools in terms of thegain scores are given in Table 1.

Table 1. Results of random effects one-way ANOVA model

Variance components	Variance	df	Variance/df	p
Student level	1247.87	45.17	27.62	0.00
School level [u_{0j}]	30.99	11.39	2.72	0.01
Variable	Estimate	S. E.	Est./S.E.	p
Intercept	5.04	1.15	4.37	0.00

According to Table 1, the variance of students' gains in student level is 1247.87, while the variance of students' gains in school level is 30.99. According to these values, the intra-class correlation coefficient is calculated as 0.02. According to this value, it can be concluded that approximately 2% of the differences in the gain scores observed among the students arise from the difference in the average gain scores between the schools and 98% of it originates from differences in student level. According to Table 1, the average gain score of students is 5.04. However, there is a significant difference between schools in terms of the gain score (p <.05).

3.3. Findings on the effect of student and school characteristics on student gain scores

The results of the first level random intercept model are given in Table 2.

Table 2. The results of the first level random intercept model

Variable	Estimate	S. E.	Est./S.E.	*p*	Pratt index
Intercept	10.83	2.50	4.32	0.00	
Gender	-1.37	1.48	-0.92	0.36	-
Socioeconomic level	-0.13	0.04	-3.02	0.01	0.09
Variance components	Variance	df	Variance/df	p	
Residual	1241.31	43.38	28.61	0.00	
School level $[u_{0j}]$	26.03	9.91	2.63	0.01	

According to Table 2, while the socioeconomic level variable has a statistically significant effect on the gain score of students (p <.05), the gender variable does not have a statistically significant effect on the gain score (p > .05). According to the results, the effect of the socioeconomic level is negative and at a quite low level. According to this, a decrease of one standard deviation in the score in socioeconomic level resulted in an increase of 0.13 in the gain score of students. Besides, when the Pratt index value of the socioeconomic level is examined, it appears that this variable has no effect on the students' gain scores in a practical sense. The results of the regression model in which means are outcomes are given in Table 3.

Table 3. The results of the regression model in which means are outcomes

Variable	Estimate	S. E.	Est./S.E.	*p*
Intercept	7.09	2.90	2.44	0.01
School size	-0.00	0.01	-1.12	0.26
School resources	-0.09	0.47	-0.18	0.86
Variance components	Variance	df	Variance/df	p
Residual	1247.86	45.16	27.63	0.00
School level $[u_{0j}]$	29.87	10.95	2.73	0.01

According to Table 3, school size and resources do not have a statistically significant effect on students' gain score (p >.05). Therefore, the second level variables which are added to the model cannot explain the variance in the gain scores observed between schools.

3.4. Findings of the models using a categorical growth model

Frequencies related to the values calculated based on schools' growth categories are given in Appendix 2, and results based on statistics are given in Figure 3.

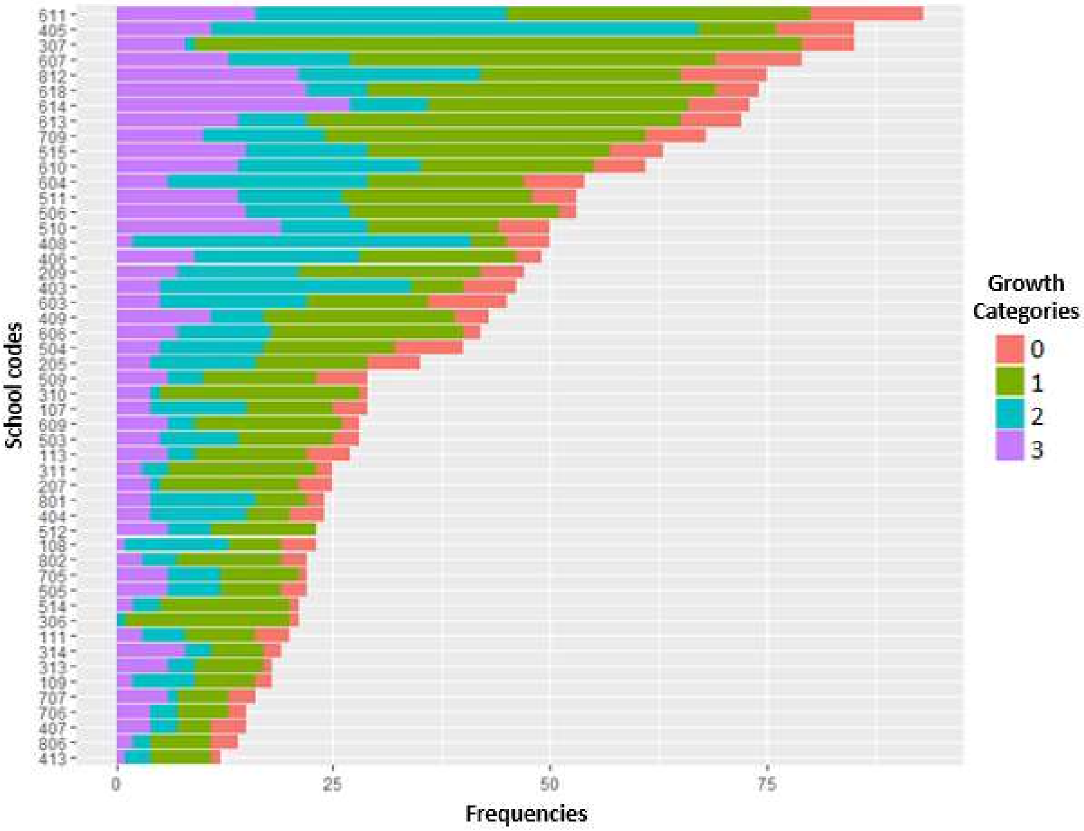


Figure 3. Frequency distribution of growth categories of schools

Upon examining schools in terms of growth categories according to Figure 3 and Appendix 2, it is seen that most students (43%) are in Category 1 and few (11.1%) are in Category 0. In other words, while most of the students are at the basic level, a small proportion of students consist of students who have downgraded their performance to a lower level. In addition, the school with the highest number of students (7%) who increased their performances is the school with the code of 614, and the lowest (0%) is the school with the code of 306. The school in which the highest number of students (5.9%) whose performance is downgraded to a lower level is the school with the code of 611 and the school with the lowest number (0%) is coded as 512.

3.5. Findings of differences between schools in terms of categorical growth model

The results regarding the models established to examine the differences between schools in terms of categorical growth model are given in Table 4. According to Table 4, the expected possibility of students to be in Category 2 than to be in Category 3 is exp{0.17}/1+ exp{0.77}+exp{0.17}+exp{-0.53}=1.18/4.92= 0.23; the expected possibility of being in Category 1 is exp {0.77}/1 +exp {0.77} + exp {0.17} + exp {-0.53} = 2.16/4.92 = 0.44; the expected possibility of being in Category 0 is exp{{-0.53}/1+ exp{0.77}+ exp{0.17}+ exp{-0.53} = 0.584.92=0.12. The possibilities of all categories except for the Category 2 are statistically significant (*p*<.05). In addition, intercept variance is significant in the first and second categories, and there are significant differences among schools in these categories (*p*<.05).

Table 4. Results of random effects one-way ANOVA model

Fixed Effects	Estimate	S. E.	t	df	*p*
Category 0 for Intercept 1 β0(0)					
Intercept 2 γ00(0)	-0.53	.09	-5.67	49	<0.001
Category 1 for Intercept 1 β0(1)					
Intercept 2 γ00(1)	0.77	0.09	8.69	49	<0.001
Category 2 for Intercept 1 β0(2)					
Intercept 2 γ00(2))	0.17	0.13	1.31	49	0.20
Variance components	Standard deviation	Variance	df	X^2	p
Intercept 1 (0), u0(0)	0.38	0.14	49	57.88	0.18
Intercept 1 (1) u0(1)	0.43	0.19	49	92.87	<0.001
Intercept 1 (2) u0(2)	0.77	0.60	49	160.16	<0.001

3.6. Findings of the effect of student and school characteristics on categorical growth

The results of the first level random intercept model are presented in Table 5. According to Table 5, female students are more than twice (*p* <.05) as likely to be in Category 1 instead of Category 3 (exp{0.79}= 2.20) than male students. However, female students are 31% less likely (*p* <.05) to be in Category 2 rather than Category 3 compared to male students (exp{-0.37}= 0.69). When the gender variable is controlled, an increase by one point in the socioeconomic level of the students is expected to increase the possibility of being in Categories 0, 1, or 2 instead of Category 3 by 0.01 points (*p* <.05). In this context, the data showed that the socioeconomic level has no effect on the students' categories in a practical sense. Another observation is that possibility of first level variables is not statistically significant in terms of other categories (*p*> .05). The results of the regression model in which means are outcomes are given in Table 6.

Table 5. The results of the first level random intercept model

Fixed Effects	Estimate	S. E.	t	df	*p*
Category 0 for Intecept 1 β0(0)					
Intecept 2 γ00(0)	-0.50	0.09	-5.48	49	<0.001
Gender β1(0), Intecept 2, γ10(0)	0.23	0.16	1.36	1848	0.17
Socioeconomic level β2(0), Intecept 2, γ20(0)	0.01	0.001	3.08	1848	.002
Category 1 for Intercept 1 β0(1)					
Intecept 2 γ00(1)	0.77	0.89	8.68	49	<0.001
Gender β1(1), Intecept 2, γ10(1)	0.79	0.13	5.79	1848	<0.001
Socioeconomic level β2(1) Intecept 2, γ20(1)	0.01	0.001	2.42	1848	.02
Category 2 for Intercept 1 β0(2)					
Intecept 2 γ00(2))	0.15	0.14	1.07	49	.29
Gender γ20(1) Intecept 2, γ10(2)	-0.37	0.16	-2.35	1848	.02
Socioeconomic level β2(2), Intecept 2, γ20(2)	0.01	0.003	3.25	1848	.001
Variance components	Standard deviation	Variance	df	X^2	p
Intercept 1 (0), u0(0)	0.33	0.11	49	53.15	0.32
Intercept 1 (1) u0(1)	0.44	0.19	49	91.14	<0.001
Intercept 1 (2) u0(2)	0.79	0.62	49	157.63	<0.001

Table 6. The results of the regression model in which means are outcomes

Fixed Effects	Estimate	S. E.	t	df	*p*
Category 0 for Intercept 1 β0(0)					
Intercept 2 γ00(0)	-0.54	0.10	-5.43	47	<0.001
School size, γ01(0)	0.0001	0.0001	1.59	47	0.12
School resources, γ02(0)	-0.002	0.04	-0.04	47	0.96
Category 1 for Intercept 1 β0(1)					
Intercept 2 γ00(1)	0.76	0.09	8.62	47	<0.001
School size, γ01(1)	-0.0001	0.0001	-0.15	47	0.88
School resources, γ02(1)	-0.05	0.04	-1.01	47	0.32
Category 2 for Intercept 1 β0(2)					
Intercept 2 γ00(2))	0.14	0.12	1.12	47	0.27
School size, γ01(2)	0.0001	0.0001	2.86	47	0.01
School resources, γ02(2)	-0.09	0.05	-1.89	47	0.06
Variance components	Standard deviation	Variance	df	X^2	p
Intercept 1 (0), u0(0)	0.39	0.15	47	56.99	0.15
Intercept 1 (1) u0(1)	0.44	0.19	47	90.55	<0.001
Intercept 1 (2) u0(2)	0.76	0.57	47	145.25	<0.001

According to Table 6, when the school resources are controlled for, an increase by one person in the size of schools would decrease the possibility of students' being in Category 2 rather than Category 3 by 0.09 (*p* <.05). Besides, the possibility of second level variables in terms of other categories is not statistically significant (*p*> .05).

4. DISCUSSION and CONCLUSION

In this study, the growth in students' comprehension skills was examined using the gain score and categorical growth models. Results showed that some students did not achieve sufficient gains to advance to higher performance levels. Although some schools' average gains were higher, their performance was still not significant enough in terms of tests' standards. Moreover, the analyses demonstrated that the student gain scores and growth categories varied significantly among the schools. In addition, the study was able to determine student and school characteristics that have an impact on the students' gain scores and categorical growth.

According to the results obtained from the gain score and categorical growth models, there was a significant difference between schools and students. When the average gain score values regarding schools and students were examined, the values proved to be positive. In this sense, it can be said that the students increased their scores from sixth to seventh grade in general. This is an expected situation upon considering the structure of comprehension skills in language (Crawford et al., 2001; Herbers et al., 2012). However, the results obtained from the gain model need to be evaluated cautiously (Betebenner & Linn, 2009; Castellano & Ho, 2013; Pike, 1991). The reason for this is to determine with the help of gain model how much the student scores have increased or decreased over one year. In other words, there is no information concerning the starting positions of the students in this model. In this case, the results of the model can be affected by the problem known as floor and ceiling effect (Rock and Pollack, 2002).

According to the results obtained from the categorical growth model, students were most likely to perform at the basic level in both the sixth and seventh grade. The lowest possibility was that students in the sixth grade performed at the proficient level, but in the seventh grade, they performed at the basic level. This can be considered as an expected result, given that the average performance of the participants in the study is low. In this sense, considering the fact that students' showing growth corresponds to improving in terms of levels in the categorical growth model (Ryser & Rambo-Hernandez, 2013), it can be stated that students within the context of the study have a low level of performance in achieving the proficient standards. Thus, although student growth has increased, this increase is not at the proficient level according the results of this study.

This study showed that gender had no effect on the significant difference in gain scores. The reason why it does not have any effect on the difference in the gain scores of students may be due to the ceiling effect; if the female students are more successful than the male students, it is expected that females' gain scores would be less. The indicators obtained from the categorical growth model revealed that female students are more likely to improve from the basic level to the proficient level than the male students. This situation between the female students and male students in terms of growth categories are in line with some studies in the literature (Denton & West, 2002; Kurdek & Sinclair, 2001). The fact that female students are more successful compared to male students also show similarity in terms of academic performance of students (Anıl, Özer – Özkan, & Demir, 2015; Büyüköztürk et al., 2014; Taş et al., 2016) in Turkey. This indicates that female students continue to display higher performance than expected at advancing grade levels than male students. This success may have resulted from the increase in the projects and programs especially intended for girls' education in recent years in Turkey.

It was determined that the socioeconomic level of the students had a significant negative effect on the gain scores, but this effect was not significant in a practical sense. At the same time, students' socioeconomic level was determined to have a significant negative effect on the possibility of passing towards from the basic level to the proficient level. This can be an example of the state of academically resilient students. Turkish researchers (e.g. Dinçer & Oral, 2010; Yavuz & Kutlu, 2016) studying academically resilient students have found that evidence to suggest that students can be academically successful in spite of the disadvantages socioeconomic level cause. Considering the participants of this study, it can be stated that the students, in general, came from a socioeconomically disadvantaged district.

The findings of this study related to school characteristics are in parallel with some studies in the literature (Glewwe, Hanushek, Humpage, &Ravina, 2011; Leithwood & Jantzi, 2009). In this context, it can be stated that the physical characteristics of schools do not play an important role in students' gain scores. It can be expressed that a similar situation can be observed in the academic achievements of students in Turkey. In addition, the fact that the study group was gathered from only one district of Ankara and that the district schools' have similar physical resources may have affected the results. It has been shown that school size and school resources which are among school characteristics, displayed no significant difference in growth categories except for the category of "remaining at the proficient level". According to the results of the analysis, the decrease in the size of the school increases the possibility of the students to remain at the proficient level. This is in line with the results of the meta-analysis study conducted by Leithwood and Jantzi (2009) on the relationship between school size and student achievement, but not in parallel with Stevenson (1996)'s study. Furthermore, given that school size has no effect on other categorical growths in this study, it can be stated that this finding is not effective in a practical sense.

It is suggested in this study that there are differences between the students and the schools in terms of their growth level. Given the different aspects gained about students' performance with these models, it is recommended to utilize different growth models in schools. In these assessments, especially because there are performance levels that have a full meaning/equivalence for the education shareholders rather than the scores of the students, the outputs of these models can be helpful for the education shareholders (Slaughter, 2008). Thus, educational stakeholders can monitor the student growth and determine whether or not the students are at the expected level of performance; thereby relevant educational institutions can take effective measures. Considering the results of this study concerning the characteristics of students and schools, for the purpose of increasing student growth, educational stakeholders may construct policies for other school characteristics that can assist in student growth.

The results of this study should be evaluated within the scope of its limitations. Since vertically scaled tests could not be used in the study and pseudo common items are used, it would be more appropriate for future researchers to use vertically scaled tests in conducting studies where students are monitored. It would be also efficient to use tests including more items and giving more information with a larger theta range. In addition, because the study is conducted in Altındağ district of Ankara, which has a low socioeconomic status and homogenous features within the district, future researchers may want to conduct a similar study in a more heterogeneous and larger study group. Aside from this, the grade level may also be included in the analysis. In future studies to be conducted, a similar study could investigate for reading comprehension, which is more general instead of language comprehension skills. In addition, a limited number of variables regarding the student and school characteristics were used in this study. For this reason, a similar study could be carried out by including other student and school characteristics that could be predictive of student growth.

Acknowledgements

This paper was extracted from the first author's Ph.D. thesis

ORCID

Hatice Cigdem YAVUZ https://orcid.org/0000-0003-2585-3686
Ömer KUTLU https://orcid.org/0000-0003-4364-5629

5. REFERENCES

Anıl, D., Özer Özkan, Y. ve Demir, E. (2015). *PISA 2012 araştırması ulusal nihai rapor*. T.C. Millî Eğitim Bakanlığı, Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü, Ankara.

Arnold, D. H., & Doctoroff, G. L. (2003). The early education of socioeconomically disadvantaged children. *Annual Review of Psychology, 54,* 517–545.

Ayral, M., Özdemir, N. ve Sadıç, Ş. (2011). *Altındağ İlçesi Öğrenme Düzeyi Araştırması Raporu*. Altındağ Rehberlik ve Araştırma Merkezi, Ankara. Retrieved February 20, 2017 from http://altindagram.meb.k12.tr/meb_iys_dosyalar/06/01/334395/dosyalar/2012_12/18012758_renmedzeyiaratrmas_1.pdf

Betebenner, D. W. (2009). *Growth, standards and accountability*. Retrieved February 25, 2017 from the National Center for the Improvement of Educational Assessment: http://www.nciea.org/publications/normative_criterion_growth_DB08.pdf

Briggs, D., & Betebenner, D. W. (2009, April). *Is growth in student achievement scale dependent*? Paper presented at the invited symposium Measuring and Evaluating. Changes in Student Achievement: A Conversation About Technical and Conceptual Issues at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Browne, W. J. & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 1*, 473–514.

Bursal, M., Buldur, S. & Dede, Y. (2015). Alt sosyo-ekonomik düzeyli ilköğretim öğrencilerinin 4-8. sınıflar fen ve matematik ders başarıları: Cinsiyet perspektifi [Science and mathematics course success of elementary students in low socio-economic status among 4th-8th grades: Gender perspective]. *Eğitim ve Bilim, 40*(179) 133-145.

Büyüköztürk, Ş., Çakan, M., Tan, Ş. ve Atar, H. Y. (2014). *TIMSS 2011 ulusal matematik ve fen raporu- 8. Sınıflar*. T.C. Millî Eğitim Bakanlığı, Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü, Ankara.

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models.* Washington, DC: Council of Chief State School Officers. Retrieved February 20, 2017 from: http://scholar.harvard.edu/files/andrewho/filesa_pracitioners_guide_to_growth_models.pdf

Chavent, M., Kuentz-Simonet, V., Labenne, A., & Saracco, J. (2014). *Multivariate analysis of mixed data: The PCAmixdata R package*. arXiv preprint arXiv:1411.4911.

Cheti, N., & Birgitta, R. (2012). *The effect of school resources ontest scores in England*, ISER Working Paper Series, No. 2012-13.

Crawford, L., Tindal, G., & Steiber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment, 7*(4), 303–323.

Denton, K., & West, J. (2002). *Children's reading and mathematics achievement in kindergarten and first grade* (NCES 2002-125). Washington, DC: National Center for Education Statistics.

Dinçer, M. A. ve Oral, I. (2010). *Türkiye'de devlet liselerinde akademik yılmazlık profili: PISA 2009 Türkiye verisinin analizi*. İstanbul: Eğitim Reformu Girişimi.

Ergin-Aydemir, S. ve Sünbül, Ö. (2016). Matematik bilişsel gelişiminin örtük büyüme modeli ile izlenmesi [Monitoring the mathematics cognitive development with the latent growth modeling]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 16*(1), 20-40.

Erman-Aslanoğlu, A. ve Kutlu, Ö. (2015). Factors related to the reading comprehension skills of 4th grade students according to data of PIRLS 2001 Turkey. *Journal of Educational Sciences Research, 5*(2), 1-18.

Glewwe, P. W., Hanushek, E. A., Humpage, S. D., & Ravina, R. (2011). *School resources andeducational outcomes in developing countries: a review of the literature from 1990 to 2010.*Working Paper 17554. Cambridge, MA: National Bureau of Economic Research.

Goldschmidt, P., Choi, K., & Beaudoin, J. P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance*. Washington, DC: Council of Chief State School Officers.

Gong, B. (2004). *Models for using student growth measures in school accountability*. Paper presented at the Council of Chief State School Officers' "Brain Trust" on Value-added Models, Washington, DC.

Gonzalez, J. & Wilberg, M. (2017). *Applying Test Equating Methods Using R*. New York: Springer International Publishing.

Güzle-Kayır, Ç. ve Erdoğan, M. (2015). The variation in Turkish students' reading skills based on PISA 2009: The effects of socio-economic and classroom-related factors. *International Online Journal of Educational Sciences, 7*(4), 80 – 96.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for the item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*, 3–24.

Hanushek, E. A. (2006). School resources. in Hanushek, E. A., and Welch, F. (eds) *Handbook of the Economics of Education*, (Amsterdam: Elsevier) 865-908

Heck, R. H. (2006). Assessing school achievement progress: Comparing alternative approaches. *Educational Administration Quarterly, 42*(5), 667-699.

Herbers, J. E., Cutuli, J. J., Supkoff, L. M., Heistad, D., Chan, C.-K., Hinz, E., & Masten, A. S. (2012). Early reading skills and academic achievement trajectories of students facing poverty, homelessness, and high residential mobility. *Educational Researcher, 41*, 366–374.

Hughes, J. N., Luo, W., Kwok, O.-M., & Loyd, L. K. (2008). Teacher-student support, effortful engagement, and achievement: A 3-year longitudinal study. *Journal of Educational Psychology, 100*(1), 1–14.

Husain, M., & Millimet, D. (2009). The mythical ‘‘boy crisis’’? *Economics of Education Review, 28*, 38-48.

Krueger, A. B., & Lindahl (2001). Education for growth: Why and for whom?, *Journal of Economic Literature, 39*(4), 1101–1136.

Kurdek, L. A., & Sinclair, R. J. (2001). Predicting reading and mathematics achievement in fourth-grade children from kindergarten readiness scores. *Journal of Educational Psychology, 93*(3), 451–455.

Kutlu, Ö., Yıldırım, Ö., Bilican, S. ve Kumandaş, H. (2011). İlköğretim 5. sınıf öğrencilerinin okuduğunu anlamada başarılı olup olmama durumlarının kestirilmesinde etkili olan değişkenlerin incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 2*(1), 132-139.

Laird, E. (2008). *Tapping into the power of longitudinal data: A guide for school leaders.* Retrieved January 21, 2018 from http://www.dataqualitycampaign.org/.

Leithwood, K., Edge, K., & Jantzi, D. (1999). *Educational accountability: The state of the art.* Gütersloh, Germany: Bertelsmann Foundation.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). *Standard setting: A Bookmark approach*. In D. R. Green (Chair), IRT-based standard setting procedures utilizing behavioral anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix AZ.

Liu, Y., Zumbo, B. D., & Wu, A. D. (2014). Relative importance of predictors in multilevel modeling. *Journal of Modern Applied Statistical Methods, 13*(1), 1-22.

Luppescu, S. (2005). Virtual equating. *Rasch Measurement Transactions, 19*(3), 10-25.

McCoach, B. D., O’Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierarchical linear model of children’s reading growth during the first 2 years of school. J*ournal of Educational Psychology, 98*, 14-28.

Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children’s growth trajectories in reading and mathematics: Who falls increasingly behind? *Journal of Learning Disabilities, 44*, 472- 488.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2001). PIRLS 2001 International Report. Chestnut Hill, MA: International Study Center. Retrieved February 20, 2017 from https://timssandpirls.bc.edu/pirls2001i/pdf/p1_IR_book.pdf.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User’s Guide* (Eighth Edition). Los Angeles, CA: Muthén & Muthén

National Assessment Agency. (2008). *14-19 Reforms*. Retrieved February 3, 2017 from: http://www.naa.org.uk/

Nayır, F. (2013). Eğitimde kalite geliştirme sürecinde okul değerlendirmenin rolü [The role of school evaluation in the process of improving quality of education]. *Celal Bayar Üniversitesi Sosyal BilimlerDergisi, 11*(2), 119-134.

Nese, J., Biancarosa, G., Anderson, D., Lai, C., Alonzo, J., & Tindal, G. (2012). Within- year oral reading fluency with CBM: A comparison of models. *Reading & Writing, 25*(4), 887-915.

No Child Left Behind Act of 2001, P. L. 107-110, 20 U.S.C. 6319 (2002).

OECD (2014a). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science* (Volume I, Revised edition, February 2014), PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264201118-en

OECD. (2014b). PISA 2012 technical report. Paris: OECD Publications.

OECD (2016). PISA 2015 Results (Volume I): Excellence and Equity in Education, PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264266490-en

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research, and Evaluation*, 7(1), 1-3.

Özer Özkan, Y. (2016). Okulları başarılarına göre sınıflandırmada etkili olan değişkenlerin PISA 2012 Türkiye verileri aracılığıyla incelenmesi [The impact of school properties to mathematics literacy in the PISA 2012 Turkey sample]. *International Online Journal of Educational Sciences, 8*(2), 117-130.

Palardy, G. J. (2008). Differential school effects among low, middle, and high social class composition schools: A multiple group, multilevel latent growth curve analysis. *School Effectiveness and School Improvement*, *19*, 21–49.

Pike, G. R. (1991). Using structural equation models with latent variables to study student growth and development. *Research in Higher Education, 32(*5), 499-524.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publision.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T., Jr. (2011). *HLM 7: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Rock, D. A., & Pollack, J. M. (2002). *A model-based approach to measuring cognitive growth in pre-reading and reading skills during the kindergarten year.* ETS Research Report Series, 2002(2).

Ryser, G., & Rambo-Hernandez, K. (2014). Using growth models to measure school performance: Implications for gifted learners. *Gifted Child Today, 37*(1), 17–23.

Shin, T., Davison, M. L., Long, J. D., Chan, C., & Heistad, D. (2013). Exploring gains in reading and mathematics achievement among regular and exceptional students using growth curve modeling. *Learning and Individual Differences, 23*, 92-100.

Skibbe, L. E., Connor, C. M., Morrison, F. J., & Jewkes, A. M. (2010). Schooling effects on preschoolers' self-regulation, early literacy, and language growth. *Early Childhood Research Quarterly, 26*, 42-49.

Slaughter, R. (2008). *Measuring middle school achievement growth with student growth percentile methodology.* NERA Conference http://digitalcommons.uconn.edu/nera_2008/20

Stevenson, H. (2006) Moving towards, into and through principalship: Developing a framework for researching the career trajectories of school leaders. *Journal of Educational Administration, 44*(4), 408-420.

Taş, U. E., Arıcı, Ö., Özarkan, H. B. ve Özgürlük, B. (2016). *PISA 2015 ulusal raporu*. T.C. Millî Eğitim Bakanlığı, Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü, Ankara.

Türkoğlu, M. E. ve Aypay, A. (2015). Özel okul öğretmenlerinin öğretmen hesap verebilirliğine dair düşünceleri [Private school teachers' thoughts about teacher accountability]. *Eğitimde Politika Analizi Dergisi, 4*(1), 7-32.

U.S. Department of Education. (2010, June 24). Race to the Top Assessment Program. Announcement of Race to the Top Program by U.S. Department of Education. Retrieved February 18, 2017 from http://www2.ed.gov/programs/racetothetop-assessment/index.html

Welch, C., Dunbar, S., & Rickels, H. (2016). *Measuring student growth in Iowa with Iowa assessments*. Retrieved February 25, 2017 from https://itp.education.uiowa.edu/ia/documents/Measuring-Student-Growth-in-Iowa-with-the-Iowa-Assessments.pdf

Yapar, T. (2014). *İngilizce okuma becerisindeki gelişimin Madde Tepki Kuramı ve örtük büyüme modellemesiyle incelenmesi* [An investigation of the development in English reading skill by using item response theory and latent growth modeling]. (Unpublished doctorate thesis). Hacettepe Üniversitesi, Ankara.

Yavuz, H. Ç., & Kutlu, Ö. (2016). Ekonomik Bakımdan Dezavantajlı Öğrencilerin Akademik Yılmazlık Düzeylerinin Bazı Koruyucu Faktörler Açısından İncelenmesi [Investigation of the factors affecting the academic resilience of economically disadvantaged high school students]. *Eğitim ve Bilim, 41*(186), 1-19.

6. APPENDICES

Appendix 1. Descriptive statistics related to the values calculated according to the gain score model of schools

School code	N	Mean	Standard Deviation	Min	Max	School code	N	Mean	Standard Deviation	Min	Max
107	29	-7.02	25.92	-57.87	40.36	705	22	2.90	31.66	-43.04	68.08
108	23	-1.15	32.03	-54.53	74.77	706	15	4.02	45.16	-54.53	84.08
109	18	6.71	35.72	-40.26	84.08	707	16	18.83	56.72	-55.35	112.18
111	20	6.15	30.01	-40.73	47.37	709	68	7.51	32.79	-64.82	83.36
113	27	3.67	33.32	-64.82	70.66	801	24	-4.98	37.64	-53.75	97.83
205	35	0.49	31.59	-55.35	67.54	802	22	6.37	39.25	-61.83	70.66
207	25	8.89	31.09	-55.35	82.10	806	14	2.65	39.59	-61.83	57.33
209	47	-0.71	32.60	-53.09	81.89	812	75	10.07	40.85	-55.35	112.18
306	21	0.29	28.38	-40.73	47.37						
307	85	0.10	32.22	-57.87	84.08						
310	29	10.39	28.33	-42.15	74.77						
311	25	17.84	33.36	-54.02	82.10						
313	18	0.09	28.78	-54.53	54.93						
314	19	18.18	38.21	-53.75	84.08						
403	46	-8.16	34.13	-59.75	67.54						
404	24	-0.97	30.17	-43.04	67.54						
405	85	-0.53	31.36	-54.02	73.42						
406	49	4.37	35.01	-56.46	128.36						
407	15	-2.51	45.49	-54.53	97.83						
408	50	-12.95	29.00	-59.75	68.35						
409	43	8.49	34.22	-55.35	70.66						
413	12	6.83	30.57	-27.53	81.89						
503	28	3.14	38.85	-57.87	96.45						
504	40	2.09	38.19	-55.35	101.52						
505	22	15.38	39.73	-54.02	70.66						
506	53	14.72	37.23	-57.87	112.18						
509	29	-0.64	39.63	-53.75	88.10						
510	50	15.32	37.12	-54.53	96.45						
511	53	17.45	44.22	-64.82	101.52						
512	23	15.42	39.86	-48.02	82.10						
514	21	-3.06	31.10	-55.35	70.66						
515	63	5.69	38.19	-57.87	84.08						
603	45	-6.01	36.61	-57.87	97.83						
604	54	-3.98	33.91	-57.87	84.08						
606	42	0.83	31.75	-64.82	61.28						
607	79	8.20	31.36	-57.87	84.08						
609	28	8.34	36.31	-55.35	112.18						
610	61	4.01	34.11	-61.83	144.89						
611	93	4.98	37.60	-59.75	101.52						
613	72	4.59	31.89	-61.83	74.77						
614	73	16.63	37.73	-48.02	112.18						
618	74	20.27	37.77	-44.51	127.46						
Total	2004	5.24	35.78	-64.82	144.89						

Appendix 2. Frequencies related to the values calculated based on schools' growth categories

School code	N	f (0)	%	f (1)	%	f (2)	%	f (3)	%	School code	N	f (0)	%	f (1)	%	f (2)	%	f (3)	%
107	29	4	1.8	10	1.2	11	2.1	4	1.0	611	93	13	5.9	35	4.1	29	5.4	16	4.1
108	23	4	1.8	6	.7	12	2.2	1	.3	613	72	7	3.2	43	5.0	8	1.5	14	3.6
109	18	2	.9	7	.8	7	1.3	2	.5	614	73	7	3.2	30	3.5	9	1.7	27	7.0
111	20	4	1.8	8	.9	5	.9	3	.8	618	74	5	2.3	40	4.6	7	1.3	22	5.7
113	27	5	2.3	13	1.5	3	.6	6	1.6	705	22	1	.5	9	1.0	6	1.1	6	1.6
205	35	6	2.7	13	1.5	12	2.2	4	1.0	706	15	2	.9	6	.7	3	.6	4	1.0
207	25	4	1.8	16	1.9	1	.2	4	1.0	707	16	3	1.4	6	.7	1	.2	6	1.6
209	47	5	2.3	21	2.4	14	2.6	7	1.8	709	68	7	3.2	37	4.3	14	2.6	10	2.6
306	21	1	.5	19	2.2	1	.2	0	0.0	801	24	2	.9	6	.7	12	2.2	4	1.0
307	85	6	2.7	70	8.1	1	.2	8	2.1	802	22	3	1.4	12	1.4	4	.7	3	.8
310	29	1	.5	23	2.7	1	.2	4	1.0	806	14	3	1.4	7	.8	2	.4	2	.5
311	25	2	.9	17	2.0	3	.6	3	.8	812	75	10	4.5	23	2.7	21	3.9	21	5.4
313	18	1	.5	8	.9	3	.6	6	1.6										
314	19	2	.9	6	.7	3	.6	8	2.1										
403	46	6	2.7	6	.7	29	5.4	5	1.3										
404	24	4	1.8	5	.6	11	2.1	4	1.0										
405	85	9	4.1	9	1.0	56	10.5	11	2.8										
406	49	3	1.4	18	2.1	19	3.6	9	2.3										
407	15	4	1.8	4	.5	3	.6	4	1.0										
408	50	5	2.3	4	.5	39	7.3	2	.5										
409	43	4	1.8	22	2.6	6	1.1	11	2.8										
413	12	1	.5	7	.8	3	.6	1	.3										
503	28	3	1.4	11	1.3	9	1.7	5	1.3										
504	40	8	3.6	15	1.7	12	2.2	5	1.3										
505	22	3	1.4	7	.8	6	1.1	6	1.6										
506	53	2	.9	24	2.8	12	2.2	15	3.9										
509	29	6	2.7	13	1.5	4	.7	6	1.6										
510	50	6	2.7	15	1.7	10	1.9	19	4.9										
511	53	5	2.3	22	2.6	12	2.2	14	3.6										
512	23	0	0.0	12	1.4	5	.9	6	1.6										
514	21	1	.5	15	1.7	3	.6	2	.5										
515	63	6	2.7	28	3.3	14	2.6	15	3.9										
603	45	9	4.1	14	1.6	17	3.2	5	1.3										
604	54	7	3.2	18	2.1	23	4.3	6	1.6										
606	42	2	.9	22	2.6	11	2.1	7	1.8										
607	79	10	4.5	42	4.9	14	2.6	13	3.4										
609	28	2	.9	17	2.0	3	.6	6	1.6										
610	61	6	2.7	20	2.3	21	3.9	14	3.6										
Total	2004	222	100	861	100	535	100	386	100										

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 506–521
https://dx.doi.org/10.21449/ijate.625423

Published at http://www.ijate.net http://submit.ijate.net/en **_Research Article_**

Mathematics Teaching Anxiety Scale: Construction, Reliability and Validity

Vesile Alkan [1,*], Tolga Coşguner [1], Yücel Fidan [1]

[1]Pamukkale University, Faculty of Education, Kınıklı Campus, 20070, Denizli, Turkey

ARTICLE HISTORY

Received: 27 June 2019
Revised: 18 August 2019
Accepted: 25 September 2019

KEYWORDS

Mathematics anxiety,
Mathematics teaching anxiety
Prospective teachers,
Reliability,
Validity,
Scale

Abstract: This study aimed to develop mathematics teaching anxiety scale for prospective primary school teachers. It was designed based on survey method and conducted with four sampling group consisting of 956 prospective primary school teachers at Education Faculties in Turkey. First sampling group was consisted of 404 prospective primary school teachers and 96 out of it were involved in the application of open-ended questions and 308 were involved in exploratory factor analysis. 305 prospective primary school teachers in the second sampling group participated in the confirmatory factor analysis, 108 prospective teachers in the third group were involved in criterion validity and 139 prospective teachers in the fourth one participated in the test-retest reliability analysis. As a result of the principal component analysis of the Mathematics Teaching Anxiety Scale (MTAS), it was found that the scale indicating single factor structure and consisting of 31 items (47.43% of the total variance). After suggested modifications, the scale MTAS was constructed with 19 items. 12 items were removed from the scale and the confirmatory factor analysis (CFA) was carried out with 19 items. According to CFA results ($0 \leq X2$ / df = $1.483 \leq 2$, RMSEA = 0.040, RMR = 0.050, AGFI = 0.908, TLI = 0.972, CFI = 0.976, IFI = 0.976, GFI = 0.928, NFI = 0.930 and RFI = 0.919), it was confirmed that the scale structure was consisting of 19 items and one dimension. The Cronbach's alpha coefficient of the final form of Mathematics Teaching Anxiety Scale was calculated as 0.93.

1. INTRODUCTION

A global improvement in Information and Communication Technology (ICT) and being an interconnected world cause differences in not only individuals' social lives but also their school lives. The rapid change in the world enables individuals to share their knowledge effortlessly and this situation results in being aware of improvements and innovations around the world. Due to these changes, the content of education in terms of disciplines and teaching strategies and styles of them are also changing (Voogt & Roblin, 2010; Trilling & Fadel, 2009).

Students of new world need to gain a set of competencies that would help them better coping with the compulsive demands of 21st century. In this sense, it could be said that mathematics is crucial for 21st century skills in that it enables to think analytically, critically and creatively which then enable to gain problem solving and reasoning skills. This means mathematics helps thinking analytically, having better problem-solving skills and having better reasoning abilities.

CONTACT: Vesile ALKAN ✉ vesile@pau.edu.tr Pamukkale University, Faculty of Education, Kınıklı Campus, 20070, Denizli, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

These skills are significant in providing individuals to find out the way of solving problems and looking for solutions in their lives. Therefore, learning and teaching mathematics in schools has become even more significant in today's world.

As emphasized by Tobias (1978) learning mathematics is intellectual but also emotional. Learning mathematics is related with how students can solve mathematical operations, how they can comprehend mathematical literacy and how they are competent in mathematics. However, it should be also noted that learning mathematics is also related with how students use their cognitive intelligences on how to succeed. On the one hand this suggests cognition and emotion are intertwisted in learning mathematics. On the other hand, even though mathematics and mathematical knowledge are used not only in schools but also regularly in everyday lives, students may avoid learning mathematics due to negative emotional reactions.

Many studies (Aiken, 1970; Alkan, 2009; 2010; 2011 & 2013; Ashcraft, 1995; Baloğlu, 1999; Bessant, 1992; Bourne, 1995; Campbell & Evans, 1997; Chipman, Krantz & Silver, 1992; Dowker, Sarkar, & Looi, 2016; Gierl & Bisanz, 1995; Hembree, 1990; Izard, 1972; Kitchens, 1995; Ma & Xu, 2004; Peker & Ertekin, 2011; Posamentier & Stepelman, 1986; Richardson, 1980; Skiba, 1990; Şahin, 2004; Tobias, 1978; Tobias, 1990; Vukovic, Kieffer, Bailey & Harari, 2013; Wu, Willcutt, Escovar & Menon; 2014; Zettle & Houghton, 1998; Zettle & Raines, 2000) indicated that some students at different grades of schools have negative attitudes towards mathematics which in turn cause feeling anxiety in mathematics. As suggested by given studies, it can be said that there is a lack in considering affective features of students in mathematics. In addition to this, it is suggested that students' anxiety in mathematics is attributable to such reasons like personality, parents, peers as well as teachers along with their teaching strategies and styles.

It can be accepted that teachers are one of the most powerful influences on students' learning of mathematics. Bandura (1993) emphasized that *"teachers' beliefs in their personal efficacy to motivate and promote learning affect the types of learning environments they create and the level of academic progress their students achieve"* (p. 117). From this point, it can be said that self-efficacy can be the predictor of teachers' effectiveness in mathematics (Hashmi & Shaikh, 2011; Swackhammer, Koellner, Basile, & Kimborough, 2009). Additionally, a wide body of studies (Alkan, 2009; 2011; Fiore, 1999; Geist, 2010; Sheilds, 2006; Sloan, 2010; Stuart, 2000) determined that teachers can cause, increase or reduce students' anxiety in mathematics at all levels of schooling on account of their attitudes and behaviours along with the teaching methods and the instructional strategies they use.

Swars, Daane & Giesen (2006) stated that there was a negative relationship between self-efficacy for teaching and mathematics anxiety. This means teacher with high level self-efficacy can convey their confidence in mathematics to students (Mji & Arigbabu, 2012) whereas those with low self-efficacy can cause students to feel negative attitudes towards mathematics. It was found in studies that teachers who are mathematics anxious fail in conveying important mathematical concepts and in allocating enough time for teaching these important concepts (Alkan, 2009; Dunkle, 2010; Fiore, 1999; Hembree, 1990 and Stuart, 2000). It can be also assumed that mathematics anxious teachers can transfer their negative attitudes in mathematics to their students.

Learning mathematics and teaching mathematics can be affected not only by the level of students' anxiety but also by the level of teachers' mathematics anxiety along with their teaching anxiety in mathematics (Alkan, 2009, 2011 and Baloğlu, 2001). The results of some studies indicated that there was a strong relation between teachers' mathematics anxiety and mathematics teaching anxiety (Bursal & Paznokas, 2006; Gresham, 2008; Swars et al., 2006). Furthermore, it was found that teachers' negative feelings and attitudes in teaching mathematics can create anxiety and increase the level of anxiety of students in mathematics (Alkan, 2009, 2011; Baloğlu, 1999; Beilock & Willingham, 2014; Finlayson, 2014; Furner & Berman, 2003; Sparks, 2011; Uusumaki & Nason, 2004; Vinson, 2001).

Mathematics teaching anxiety can be define as teachers' feeling negative reaction to mathematics, feeling under pressure to teach mathematics and being frustrated with the lack of progress in mathematics. Teachers who feel anxiety in teaching mathematics might have fear of explaining concepts, formulae and operations in mathematics. However, it should be noted that mathematics is cumulative; there is a relation between prior knowledge, current and further knowledge in mathematics. This means the teacher needs to clarify each topic in mathematics in order not to cause students to fall behind. In addition to this, the teacher needs to help students to comprehend each concepts and operations in mathematics clearly.

Ölmez and Cohen (2018) emphasized that teachers are expected to provide supportive classroom setting in which lessening students' negative feelings towards mathematics. Furthermore, teachers are expected to enhance students' involvement in mathematics by helping to build connections with real-life situations and also building their self-confidence in mathematics. Although these expectations are specified, it should be considered that teachers having negative attitudes towards mathematics and teaching mathematics can fail in meeting these. Therefore, it is crucial to find out the level of mathematics teaching anxiety of teachers to deal with their anxieties in teaching mathematics.

As given in many studies above, there is an association between students' negative feelings in mathematics and teachers' anxiety and teaching anxiety in mathematics. It should be noted that feeling anxiety in mathematics can be started at primary school and raise at other levels of schooling and can transfer to the professional life. Like teachers, prospective teachers' teaching efficacy and self-confidence in mathematics can have an impact on their learning mathematics and then their teaching process (Hudson, Kloosterman& Galindo, 2012). Levine (1993; 1996) claimed that prospective teachers have difficulties in teaching mathematics due to their teaching anxiety. Hence, mathematics anxious prospective teachers may avoid mathematics and mathematics related courses which in turn cause teaching in a way that unconsciously leading their students to feel anxiety in mathematics.

Prospective teachers especially for primary schools are significant resources for future mathematics lessons in schools and for improving future students' self-efficacy in mathematics. For this reason, it is needed to improve their teaching efficacy in mathematics in order to help these future teachers to be successful in their teaching in mathematics (Ryang, 2012). Gurin and et al, (2017) stated that there was a slight increase on studies conducted to find out the relation between teachers' mathematics anxiety and students' mathematic anxiety. Moreover, it is seen that there is a few studies focusing on prospective teachers' teaching anxiety in mathematics. These situations show that there is a need to investigate teachers' and prospective teachers' mathematics teaching anxiety in order to find out the ways of diminishing their and students' anxiety in mathematics. It is assumed that the results of studies focusing on mathematics teaching anxiety can contribute to the area of teaching mathematics. On the other hand, there is also need to find out the level of prospective teachers' mathematics teaching anxiety in order to help them to reduce or overcome this anxiety. Consequently, this study aimed to develop a scale for mathematics teaching anxiety based on prospective primary school teachers' perceptions.

2. METHOD

This study was designed in terms of quantitative approach to construct a scale for mathematics teaching anxiety for prospective teachers. To this view, a scale development steps were used.

2.1. Sampling

The participants of this study consisted of 956 prospective primary school teachers at Education Faculties in Turkey. These participants were included in four different sampling groups. The first group of this study was consisted of 404 prospective primary school teachers and 96 prospective teachers from this group were used in the application of open-ended questions and 308 of them ($\bar{X}$= 21.87, Sd = 1.83; female = 233, male = 75) were used for exploratory factor analysis. A total of 305 ($\bar{X}$= 21.95, Sd = 1.31; Female = 234, Male = 71) prospective primary school teachers in the second sampling group were used for confirmatory factor analysis, 108 prospective primary school teachers in the third group ($\bar{X}$= 21.80, Sd = 1.01; Female = 91, Male = 10) were used for criterion validity studies. Lastly, 139 prospective primary school teachers in the fourth sampling group (female = 111; male = 28) were included in test-retest reliability studies.

2.2. Assessment Measures

During the development of the Turkish version of Mathematics Teaching Anxiety Scale (MTAS), the steps proposed by De Vellis (2014), Tavşancıl (2006) and Erkuş (2014) were followed. In order to develop the scale, first of all, the literature and assessment tools were reviewed and examined. After that, the form including open-ended questions was given to prospective primary school teachers and based on their answers 57 items were prepared for the scale within the conceptual frame. Then, the draft scale form was sent to the experts who worked on such topics as mathematics teaching, anxiety and mathematics anxiety. This supported the content-related validity of the scale. In line with the recommendations of these experts, 5 items were removed from the form and suggested corrections were done. After the scale's items were clarified according to the views, the original form of the scale consisting of 52 items was designed.

Items were rated on a 5-point Likert type ranging from 1 to 5. The ranges of the scale were 1 (Strongly disagree), 2 (Slightly agree), 3 (Partially agree), 4 (Mostly agree), and 5 (Completely agree). Volunteer prospective primary school teachers were involved in data collection process. Before the data collection the participants were informed about the study and the data collection tool.

In order to perform confirmatory factor analysis, the Mathematics Teaching Efficacy Beliefs Instrument (MTEBI) was used. This instrument was used to measure prospective teachers' efficacy beliefs in teaching mathematics. The original scale was developed by Enochs, Smith & Huinker (2000). Its first adaptation to Turkish was carried out by Çakıroğlu (2000), and the second one was by Hacıömeroğlu & Şahin - Taşkın (2010). The current adapted version of the scale was used in the present study. This instrument was consisted of 17 items and 7 out of these items were scored reversely.

2.3. Data Analysis

SPSS 22.00 package program and AMOS 18.00 program were used to analyse the data. The principal component analysis within the scope of exploratory factor analysis (EFA) was performed using the Kaiser Criteria (eigenvalue> 1). After finding by the exploratory factor analysis that the scale was uni-dimensional, the Cronbach Alpha coefficient was calculated to determine the internal consistency of the scale. Confirmatory factor analysis was done with the help of AMOS 18.00 program (Byrne, 2009). For the criterion validity of the scale, Pearson product moment correlation coefficient was measured between the Mathematics Teaching Efficacy Belief Instrument (MTEBI) and the scale. In the analysis phase, whether the data had a univariate normal distribution in each study group was examined at first. It was determined that the data obtained from all study groups had a univariate normal distribution and the skewness and kurtosis values were between -1 and +1 (Muthén & Kaplan, 1985).

3. RESULTS

3.1. Exploratory Factor Analysis

While doing the Exploratory Factor Analysis (EFA), primarily the data gathered from the study group with whom MTAS consisting of 52 items applied was investigated. In this context, the chi-square value of the Bartlett Sphericity Test was found to be significant with 8973.88 (*p* <0.000), and the Kaiser-Meyer-Olkin value (0.949) was found to be sufficient. In the light of these results, it was determined that the data obtained from the first study group was suitable for factor analysis (Albayrak, 2006; Şencan, 2005). In order to determine the factor structure of the MTAS, a single-factor structure consisting of 31 items was determined as a result of the principal components analysis carried out based on the criteria of screen-plot and eigenvalue> 1.0 and it was revealed that this structure explained 47.43% of the total variance (Kline, 1994).

The Cronbach Alpha coefficient was preferred in the calculation of the reliability coefficient of the MTAS, since it yielded consistent results in determining the reliability of the assessment tools with a single factor structure (Tan, 2009). In this respect, the lowest acceptable value for Chronbach Alpha coefficient was determined to be ≥ 0.70. The reliability value of the MTAS was found to be 0.96, which is a high value (Hair, Anderson, Tatham & Black, 1998; Nunnally & Bernstein, 1994). The test-retest reliability coefficient of the MTAS was calculated to be 0.703 and this value was considered equal to the acceptable limit value. The factor loadings of the items on the MTAS, common variance and Cronbach Alpha coefficient for the single-factor structure of the scale is given in Table 1.

Table 1. *Results of the Exploratory Factor Analysis of Mathematics Teaching Anxiety Scale (N = 308)*

Item No	Item	Factor 1	Common Variance
M29	When a student does not understand mathematical operations, I get anxious about how to explain them. *Matematiksel işlemleri öğrenci anlamadığında nasıl açıklayacağım endişesi yaşarım.*	0.770	0.613
M27	A rise in the level differences among my students while teaching mathematics worries me. *Matematik dersini işlerken öğrencilerim arasında düzey farklılıklarının artmasından endişelenirim.*	0.747	0.591
M40	Until I gain experience in teaching, I feel fear about my lack of conveying mathematical concepts on time. *Deneyim kazanana kadar matematik kavramlarını zamanında kazandıramamaktan korkarım.*	0.737	0.631
M44	I feel worry about not being able to teach in mathematics according to my students' level. *Matematik dersini öğrencilerimin düzeylerine göre anlatamayacağım endişesi yaşarım.*	0.732	0.539
M23	The thought that I cannot concretize the abstract concepts in mathematics frightens me. *Matematik dersinde soyut kavramları somutlaştıramama düşüncesi beni korkutur.*	0.730	0.599
M35	I feel anxious while considering students' individual differences in teaching mathematics. *Matematik öğretirken bireysel farklılıkları göz önünde bulundurma zorunluluğu beni endişelendirir.*	0.729	0.646
M43	I feel worry that I do not know how to teach mathematical concepts to students. *Matematik kavramlarını kazandırırken nasıl öğreteceğimi bilmediğim için tedirgin olurum.*	0.726	0.605
M46	I feel anxious that I may fail in bringing my students having different readiness levels to the same level in mathematics. *Matematik dersinde hazırbulunuşluk düzeyi farklı olan öğrencilerimi aynı düzeye getiremeyeceğim endişesi yaşarım.*	0.724	0.559
M26	I feel anxious about not relating the content of mathematics with students' daily lives. *Matematik dersinde işlenecek konuyu günlük yaşamla ilişkilendiremeyeceğim endişesi yaşarım.*	0.722	0.635

Table 1. *Continues*

Item No	Item	Factor 1	Common Variance
M31	I feel anxious that I cannot finish the outcomes of the mathematics curriculum on time. *Matematik programındaki kazanımları zamanında bitiremeyeceğim endişesi yaşarım.*	0.722	0.525
M50	I'm afraid of losing my classroom control if I cannot solve the problems in mathematics. *Matematik dersinde problemleri çözemezsem sınıftaki hâkimiyetimi kaybetmekten korkarım.*	0.719	0.752
M52	I feel anxious about how I'm going to teach the subjects that I feel incompetent in mathematics. *Matematik dersinde kendimi yeterli hissetmediğim konuları öğrencilerime nasıl kazandıracağım endişesi yaşarım.*	0.705	0.587
M25	I'm worried about not using the appropriate method and technique in mathematics. *Matematik dersine uygun yöntem ve tekniği kullanamama endişesi yaşarım.*	0.697	0.589
M41	I'm worried about not enabling my students' to engage in mathematics actively. *Öğrencilerimin matematik dersine aktif katılımını sağlayamama endişesi yaşarım.*	0.696	0.630
M22	The thought that the student cannot comprehend when I turn a concept into a mathematical sentence (e.g. 2 + 3) makes me anxious. *Bir kavramı matematiksel cümleye (ör: 2+3) dönüştürdüğümde öğrencinin anlayamayacağı düşüncesi beni tedirgin eder.*	0.696	0.586
M34	I get anxious about designing activities that are appropriate for my students' level in mathematics. *Matematik dersinde öğrencilerimin düzeyine uygun etkinlik hazırlama endişesi yaşarım.*	0.695	0.554
M49	The thought that the level differences of the students in mathematics may reduce the interest of attending the lesson disturbs me. *Matematik dersinde öğrencilerin düzey farklılıklarının derse olan ilgiyi azaltacağı düşüncesi beni rahatsız eder.*	0.692	0.561
M39	I feel uneasy with the thought that I cannot enable my students to like mathematics. *Matematiği sevdiremeyeceğim düşüncesi beni huzursuz eder.*	0.679	0.671
M18	I am afraid that the level differences of the students in mathematics may affect my teaching pace. *Matematik dersinde öğrencilerin düzey farklılıklarının ders işleme hızımı etkilemesinden korkarım.*	0.674	0.558
M21	I am afraid that students with fewer interests in mathematics may reduce the interest of other students. *Matematik dersine ilgisi az olan öğrencilerin diğer öğrencilerin ilgisini azaltmasından korkarım.*	0.674	0.608
M37	I am afraid that families will criticize me if I cannot catch up with the mathematics curriculum. *Matematik programını yetiştiremezsem ailelerin beni eleştirmesinden huzursuz olurum.*	0.672	0.495
M33	I am afraid that school administrators will criticize me if I cannot catch up with the mathematics curriculum. *Matematik programını yetiştiremezsem okul yöneticilerinin beni eleştirmesinden korkarım.*	0.659	0.509
M19	The fact that my students have different readiness levels in mathematics frightens me in the early years of my professional life. *Meslek yaşantımın ilk yıllarında öğrencilerimin matematik dersindeki hazırbulunuşluk düzeylerinin farklı olması beni korkutur.*	0.656	0.621
M28	I am anxious since I believe that I do not have sufficient knowledge about teaching mathematics. *Matematik öğretimine yönelik yeterli bilgiye sahip olmadığımı düşündüğümden endişelenirim.*	0.651	0.634
M48	I feel fear of being humiliated by the students if I cannot solve problems in mathematics. *Matematik dersinde problemleri çözemezsem öğrencilerin gözünde küçük düşmekten korkarım.*	0.648	0.636
M24	It makes me uncomfortable to know that the next lesson I will teach is mathematics. *İşleyeceğim bir sonraki dersin matematik olduğunu bilmek beni huzursuz eder.*	0.647	0.693

Table 1. *Continues*

M15	I feel anxious if the differences in the level of the students in mathematics affect my classroom management. *Öğrencilerin matematik dersindeki düzey farklılıklarının sınıf hâkimiyetimi etkilemesinden endişelenirim.*	0.630	0.599
M30	I feel insecure about the thought that my students having level differences in mathematics can isolate themselves from the class eventually. *Matematik dersinde düzey farklılıkları olan öğrencilerimin zamanla kendilerini sınıftan soyutlayabilecekleri düşüncesi beni huzursuz eder.*	0.606	0.615
M3	I'm worried that I cannot motivate the students due to my prejudices against mathematics. *Matematiğe yönelik önyargılarımdan dolayı öğrencileri motive edemeyeceğim endişesi yaşarım.*	0.585	0.455
M13	I feel uncomfortable in mathematics since I do not have enough experience. *Yeterli deneyime sahip olmadığım için matematik dersinde kendimi huzursuz hissederim.*	0.585	0.414

Chronbach's Alpha Coefficient: 0.96

3.2. Confirmatory Factor Analysis

AMOS 18.00 program was used in order to perform the confirmatory factor analysis (CFA) of the MTAS and maximum likelihood method was opted for the estimation of model parameters (Tezbaşaran, 1997). The structure consisting of 31 items and one dimension as a result of exploratory factor analysis was tested via confirmatory factor analysis. The result of the analysis indicated that some of the items exhibited a high correlation with each other.

In this respect, the items exhibiting correlations were removed from the scale. Yet, after suggested modifications, the scale MTAS was constructed with 19 items and one dimension. The confirmatory factor analysis values of the MTAS and the suggested are illustrated in Figure 1.

It is stated that there are three types of fit that are practical for all fit measures and can be represented as absolute, incremental and restricted fit in the CFA (Schumacker and Lomax, 2010). In this study, X^2, RMSEA, GFI and RMR were used to evaluate the absolute fit. AGFI, NFI, TLI, CFI, RFI and IFI were used as incremental fit measures. The fit values for CFA are shown in Table 2.

Table 2. *Goodness of Fit Indices in the Confirmatory Factor Analysis*

X^2	X^2/df	p-value	RMSEA	GFI	RMR	AGFI	NFI	TLI	CFI	RFI	IFI
220.963*	1.483	0.000	0.040	0.928	0.050	0.908	0.930	0.972	0.976	0.919	0.976

* p<0.01

When Table 2 is examined, it is seen that X^2 value (X^2 = 220.963; df = 126, p <0.01) is significant (Timm, 2002). However, this statistic is considered to be a weak absolute fit (Timm, 2002). When the relevant literature is reviewed, it is observed that X^2 value is significant in large samples (Byrne, 1989). For this reason, X^2/df, which is another proposed statistic, was calculated and it was found that this statistic ($0 \leq X^2/df = 1.483 \leq 2$) showed good fit (Kline, 2011; Sümer, 2000). When the other fit indices were examined, it was observed that RMSEA (0.040), RMR (0.050), AGFI (0.908), TLI (0.972), CFI (0.976) and IFI (0.976) showed a good fit. The indices with acceptable fit values included GFI (0.928), NFI (0.930) and RFI (0.919) (Hair, Black, Babin & Anderson, 2014; Browne & Cudeck, 1993; Baumgartner & Homburg, 1996; Bentler, 1980; Bentler & Bonett, 1980; Marsh, Hau, Artelt, Baumert & Peschar, 2006; Schermelleh–Engel & Moosbrugger, 2003; Kline,1991). When these values are examined, it can be stated that the MTAS has a good fit. Table 3 shows the 19-item MTAS, standardized factor loadings and standard error values of this scale.

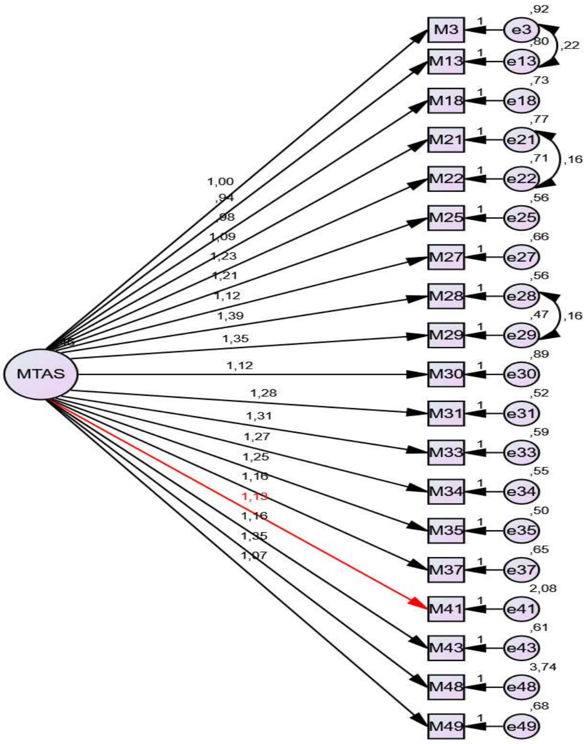


Figure 1. Results of the Confirmatory Factor Analysis of MTAS

Table 3. *Confirmatory Factor Analysis Item Statistics*

Item No		Standardized Factor Loadings	S.E.
M3	I'm worried that I cannot motivate the students due to my prejudices against mathematics. *Matematiğe yönelik önyargılarımdan dolayı öğrencileri motive edemeyeceğim endişesi yaşarım.*	0.577	
M13	I feel uncomfortable in mathematics since I do not have enough experience. *Yeterli deneyime sahip olmadığım için matematik dersinde kendimi huzursuz hissederim.*	0.580	0.096
M18	I am afraid that the level differences of the students in mathematics may affect my teaching pace. *Matematik dersinde öğrencilerin düzey farklılıklarının ders işleme hızımı etkilemesinden korkarım.*	0.614	0.111
M21	I am afraid that students with fewer interests in mathematics may reduce the interest of other students. *Matematik dersine ilgisi az olan öğrencilerin diğer öğrencilerin ilgisini azaltmasından korkarım.*	0.647	0.119
M22	The thought that the student cannot comprehend when I turn a concept into a mathematical sentence (e.g. 2 + 3) makes me anxious. *Bir kavramı matematiksel cümleye (ör: 2+3) dönüştürdüğümde öğrencinin anlayamayacağı düşüncesi beni tedirgin eder.*	0.705	0.127
M25	I'm worried about not using the appropriate method and technique in mathematics. *Matematik dersine uygun yöntem ve tekniği kullanamama endişesi yaşarım.*	0.740	0.120
M27	A rise in the level differences among my students while teaching mathematics worries me. *Matematik dersini işlerken öğrencilerim arasında düzey farklılıklarının artmasından endişelenirim.*	0.684	0.117

Table 3. *Continues*

Item No		Standardized Factor Loadings	S.E.
M28	I am anxious since I believe that I do not have sufficient knowledge about teaching mathematics. *Matematik öğretimine yönelik yeterli bilgiye sahip olmadığımı düşündüğümden endişelenirim.*	0.783	0.134
M29	When a student does not understand mathematical operations, I get anxious about how to explain them. *Matematiksel işlemleri öğrenci anlamadığında nasıl açıklayacağım endişesi yaşarım.*	0.800	0.129
M30	I feel insecure about the thought that my students having level differences in mathematics can isolate themselves from the class eventually. *Matematik dersinde düzey farklılıkları olan öğrencilerimin zamanla kendilerini sınıftan soyutlayabilecekleri düşüncesi beni huzursuz eder.*	0.628	0.125
M31	I feel anxious that I cannot finish the outcomes of the mathematics curriculum on time. *Matematik programındaki kazanımları zamanında bitiremeyeceğim endişesi yaşarım.*	0.770	0.124
M33	I am afraid that school administrators will criticize me if I cannot catch up with the mathematics curriculum. *Matematik programını yetiştiremezsem okul yöneticilerinin beni eleştirmesinden korkarım*	0.757	0.129
M34	I get anxious about designing activities that are appropriate for my students' level in mathematics. *Matematik dersinde öğrencilerimin düzeyine uygun etkinlik hazırlama endişesi yaşarım.*	0.757	0.125
M35	I feel anxious while considering students' individual differences in teaching mathematics. *Matematik öğretirken bireysel farklılıkları göz önünde bulundurma zorunluluğu beni endişelendirir.*	0.767	0.122
M37	I am afraid that families will criticize me if I cannot catch up with the mathematics curriculum. *Matematik programını yetiştiremezsem ailelerin beni eleştirmesinden huzursuz olurum.*	0.701	0.120
M41	I'm worried about not enabling my students' to engage in mathematics actively. *Öğrencilerimin matematik dersine aktif katılımını sağlayamama endişesi yaşarım.*	0.469	0.157
M43	I feel worry that I do not know how to teach mathematical concepts to students. *Matematik kavramlarını kazandırırken nasıl öğreteceğimi bilmediğim için tedirgin olurum.*	0.709	0.119
M48	I feel fear of being humiliated by the students if I cannot solve problems in mathematics. *Matematik dersinde problemleri çözemezsem öğrencilerin gözünde küçük düşmekten korkarım.*	0.427	0.203
M49	The thought that the level differences of the students in mathematics may reduce the interest of attending the lesson disturbs me. *Matematik dersinde öğrencilerin düzey farklılıklarının derse olan ilgiyi azaltacağı düşüncesi beni rahatsız eder.*	0.660	0.115

Chronbach's Alpha: 0.93

3.3. Criterion Validity

Within the scope of the criterion validity studies of the MTAS, prospective primary school teachers in the third group were asked to fill in the Mathematics Teaching Efficacy Belief Instrument and the Mathematics Teaching Anxiety Scale in order to measure the Pearson product moment correlation coefficient. It was found that the correlation coefficient showed a moderately negative (r = –0.43) and significant (p <0.01, n = 108) relationship (Büyüköztürk, 2012; Field, 2009). In the light of these results, it can be said that the MTAS has concurrent validity.

4. CONCLUSION

This study aimed to develop and examine a scale for measuring mathematics teaching anxiety (MTAS) for prospective primary school teachers. To this aim, 956 prospective primary school teachers were involved in this study in order to construct and to prove the validity and reliability of the scale. At the beginning of the study, a scale was designed with 57 items and sent to experts for content-related validity. After their judgements, the scale was structured with 52 items.

Before the factor analysis process, it was found that the chi-square value of the Bartlett Sphericity Test was significant with 8973.88 (*p* <0.000), and the Kaiser-Meyer-Olkin value was sufficient (0.949). According to the results of the exploratory factor analysis, it was found that the scale indicates single factor structure and consisting of 31 items. The reliability value of the scale with 31 items was found to be 0.96, which is a high value. In addition to this, the test-retest reliability coefficient was calculated to be 0,703 was considered equal to the acceptable limit value.

Confirmatory factor analysis was also used to determine the correlations among items. In this analysis, it was found that some items were exhibiting high correlations; therefore, those items were removed from the scale. As a result, the structure of the scale was constructed with 19 items. In terms of CFA results (0≤X2 / df = 1.483≤2, RMSEA = 0.040, RMR = 0.050, AGFI = 0.908, TLI = 0.972, CFI = 0.976, IFI = 0.976, GFI = 0.928, NFI = 0.930 and RFI = 0.919), it was confirmed that the scale structure was consisting of 19 items and one dimension. Thereafter, the criterion validity was measured and found that the scale has concurrent validity.

In conclusion, the final form of Mathematics Teaching Anxiety Scale (MTAS) for prospective primary school teachers was consisting of 19 items and the Cronbach's alpha coefficient of this scale was 0.93. It is believed that this MTAS can contribute to the area by helping to measure the level of prospective teachers' mathematics teaching anxiety. Furthermore, this scale could be one of the measurements in the area which can help other research to construct new scales and to focus on mathematics teaching anxiety in various ways.

ORCID

Vesile ALKAN https://orcid.org/0000-0002-8630-3357

5. REFERENCES

Aiken, L. R. (1970). Attitudes toward mathematics. *Review of Educational Research, 40* (4), 551–596.

Albayrak, A. S. (2006). *Uygulamalı Çok Değişkenli İstatistik Teknikleri*. Ankara: Asil Yayın Dağıtım.

Alkan, V. (2009). *The Relationship between teaching strategies and styles and pupils' anxiety in mathematics at primary schools in Turkey*. Unpublished PhD Thesis. The University of Nottingham.

Alkan, V. (2010). Matematikten nefret ediyorum! [I hate Mathematics!]. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi [Pamukkale University Journal of Education*], *28* (II), 189-199.

Alkan, V. (2011). Etkili matematik öğretiminin gerçekleştirilmesindeki engellerden biri: kaygı ve nedenleri [One of the barriers to providing effective mathematics teaching: anxiety and its causes]. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi [Pamukkale University Journal of Education], 29*(I), 89-107.

Alkan, V. (2013). Reducing mathematics anxiety: The ways implemented by teachers at primary schools in Turkey. *International J. Soc. Sci. & Education*, *3* (3), 795-807.

Ashcraft, M. H. (1995). Cognitive psychology and simple arithmetic: A review and summary of new directions. *Mathematical Cognition*, *1*, 3–34.

Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science*, *11*(5), 181–185.

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modelling in marketing and consumer research: A review. *International Journal of Research in Marketing, 13*(2), 139–161.

Baloğlu, M. (1999). A comparison of mathematics anxiety and statistics anxiety in relation to general anxiety. *Eric Document* Number (ED 436 703).

Baloglu, M. (2001). Matematik Korkusunu Yenmek. *Kuram ve Uygulamada Eğitim Bilimleri Dergisi*, *1* (1), 59–76.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist*. 28, 117-148.

Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences*, *107* (5), 1860–1863.

Beilock, S. L. & Willingham, D. T. (2014). Ask the cognitive scientist. Math anxiety: Can teachers help students reduce it? *American Educator*, *38*(2), 28-43

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

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

Bessant, K. C. (1992). Instructional design and the development of statistical literacy. *Teaching Sociology, 20*, 143–149.

Bowd, A. & Brady, P. (2003). Gender differences in mathematics anxiety among preservice teachers and perceptions of their elementary and secondary school experience with mathematics. *The Alberta Journal of Educational Research, XLIX* (1)*,* 24–36.

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

Bursal M., & Paznokas, L. (2006). Mathematics anxiety and pre-service elementary teachers' confidence to teach mathematics and science. *School Science and Mathematics*, *106* (4) 173–179.

Bourne, E. (1995). *The Anxiety and Phobia Workbook* (2nd Ed.). Akland, CA: New Harbiner Publications.

Byrne, B. M. (1989). A Primer of LISREL: Basic Applications and Programming for Confirmatory Factor Analytic Models. New York: Springer-Verlag.

Büyüköztürk, Ş. (2012). *Sosyal Bilimler İçin Veri Analizi El Kitabı, İstatistik, Araştırma Deseni SPSS Uygulamaları ve Yorumu*. Ankara: PegemA Yayıncılık.

Byrne, B. M. (2009). *Structural equation modelling with Amos: Basic concepts, applications and programming* (2nd Ed.). Mahwah, NJ: Erlbaum.

Campbell, K. & Evans, C. (1997). Gender issues in the classroom: A comparison of mathematics anxiety. *Education, 117* (3), 332–339.

Chipman, S.F., Krantz, D. H. & Silver R. (1992). Mathematics anxiety and science careers among able college women. *Psychological Science,* 3, 292–295.

Çakıroğlu, E., 2000. Preservice Elementary Teachers' Sense of Efficacy in Reform Oriented Mathematics. Yayınlanmamış Doktora Tezi, Indiana University. Retrieved from ProQuest Dissertations & Theses Global. (Order No. 9980980)

Deniz, L. ve Üldaş, İ. (2008). Öğretmen ve öğretmen adaylarina yönelik matematik kaygı ölçeği'nin geçerlik, güvenirlik çalışması. [Validity and reliability study of the mathematics anxiety scale involving teachers and prospective teachers]. *Eurasian Journal of Educational Research*, 30, 49–62.

DeVellis, R. F. (2014). *Ölçek Geliştirme Kuram ve Uygulamalar* (3. Baskı). (T. Totan, Çev. Ed.). Ankara: Nobel Akademi Yayınları.

Dowker, A., Sarkar, A. & Looi, C. Y. (2016). Mathematics anxiety: What have we learned in 60 years? *Front. Psychol.* 7, 508. DOI: 10.3389/fpsyg.2016.00508

Dunkle, S. M. (2010). Remediation of math anxiety in preservice elementary school teachers (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (Order No. 3411597)

Enochs, L., Smith, P. L., & Huinker, D. (2000). Establishing factorial validity of the mathematics teaching efficacy beliefs instrument. *School Science and Mathematics, 100*(4), 194–202.

Erkuş, A. (2014). *Psikolojide Ölçme ve Ölçek Geliştirme - I Temel Kavramlar ve İşlemler* (2.Basım). Ankara: Pegem Akademi.

Field, A. (2009). Discovering Statistics Using SPSS (3rd Ed.). London: Sage Publications Ltd.

Finlayson, M. (2014). Addressing math anxiety in the classroom. *Improving Schools, 17*(1), 99-115, DOI: 10.1177/1365480214521457

Fiore, G. 1999. Math-abused students: Are we prepared to teach them? *The Mathematics Teacher,* 92(5): 403−409.

Furner, J.M., & Berman, B.T. (2003). Math Anxiety: Overcoming a major obstacle to the improvement of student math performance. *Childhood Education*, *79*(3), 170-175.

Gardner, L. & Leak, G. (1994). Characteristics and correlates of teaching anxiety among college psychology teachers. *Teaching of Psychology, 21*(1), 28–32.

Geist, E. (2010). The anti-anxiety curriculum: Combating math anxiety in the classroom. *Journal of Instructional Psychology*, *37* (1), 24–31.

Gierl, M. J. & Bisanz, J. (1995). Anxieties and attitudes related to math in grades 3 and 6. *Journal of Experimental Education, 63* (2), 139–158.

Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., & Hall, N. C. (2013). Do girls really experience more anxiety in mathematics? *Psychological Science*, *24* (10), 2079–2087.

Gresham, G. (2008). Mathematics anxiety and mathematics teacher efficacy in elementary pre-service teachers. *Teaching Education*, *19* (3), 171–184.

Gurin, A., Jeanneret, G., Pearson, M., Pulley, M., Salinas, A. & Castillo-Garsow, C. (2017). The Dynamics of math anxiety as it is transferred through peer and teacher interactions (Technical Report MTBI-14-05M). Retrieved from Arizona State University Mathematical and Theoretical Biology Institute (MTBI) website: https://mtbi.asu.edu/sites/default/files/manuscript_0.pdf

Hacıömeroğlu, G. & Şahin–Taşkın, Ç. (2010). Sınıf Öğretmeni adaylarının matematik öğretimi yeterlik inançları. [Elementary preservice teachers' mathematics teaching efficacy belief]. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi [Uludağ University Educaton Faculty Journal], 23*(2), 539–555.

Hackett, G., & Betz, N. E. (1989). An exploration of the mathematics self-efficacy/ mathematics performance correspondence. *Journal for Research in Mathematics Education*, *20* (3), 261–273.

Hadfield, O. D. & McNeil, K. (1994). The relationship between Myers–Briggs personality type and mathematics anxiety among preservice elementary teachers. *Journal of Instructional Psychology, 21*(4), 375–384.

Hair Jr., J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate Data Analysis* (5th Ed.). Upper Saddle River, NJ: Prentice Hall.

Hair, Jr. J. F., Black, W. C., Babin, B. J. & Anderson, R. E. (2014). *Multivariate Data Analysis* (7th Ed.). USA: Pearson Education Limited

Hashmi, M. & Shaikh, F. (2011). Comparative analysis of the effect of teacher education on motivation, commitment, and self-efficacy. *New Horizons, 10*(2), 54-58.

Hembree, R. (1990). The nature, effects and relief of mathematics anxiety. *Journal of Research in Mathematics Education, 21*, 33–46.

Hudson, R., Kloosterman, P. & Galindo, E. (2012). Assessing preservice teachers' beliefs about the teaching and learning of mathematics and science. School Science and Mathematics, 112(7), 433–442.

Izard, C. E. (1972). *Patterns of Emotions: A New Analysis of Anxiety and Depression.* New York: Academic Press.

Kesici, S., & Erdoğan, A. (2009). Predicting college students' mathematics anxiety by motivational beliefs and self-regulated learning strategies. *College Student Journal*, *43* (2), 631–642.

Kitchens, A. (1995). *Defeating Math Anxiety.* Chicago: Irwin Career Education Division.

Kline, R. B. (1991). Latent variable path analysis in clinical research: a beginner's tour guide. *Journal of Clinical Psychology, 47*, 471–484.

Kline P. (1994). *An Easy Guide to Factor Analysis.* London: Routledge.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modelling.* New York: The Guilford Press.

Levine, G. (1993). Prior mathematics history, anticipated mathematics teaching style, and anxiety for teaching mathematics among pre-service elementary school teachers. Paper presented at the Annual Meeting of the International Group for Psychology of Mathematics Education, North American Chapter. (ERIC Document Reproduction Service No. ED373972).

Levine, G. (1996). Variability in anxiety for teaching mathematics among pre-service elementary school teachers enrolled in a mathematics course. Paper presented at the Annual Meeting of the American Educational Research Assocation in New York. (ERIC Document Reproduction Service No. ED398067).

Ma, X. & Xu, J. (2004). The causal ordering of mathematics anxiety and mathematics achievement: a longitudinal panel analysis. *Journal of Adolescence*, *27*(2), 165-179.

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

Mji, A. & Arigbabu, A. A. (2012). Relationships between and among pre-service mathematics teachers' conceptions, efficacy beliefs and anxiety. *International Journal of Education of Science*, 4(3), 261-270.

Muthe´n B. & Kaplan, D. A. (1985). Comparison of methodologies for the factor analysis of non–normal Likert variables. *British Journal of Mathematical and Statistical Psychology,* 38, 171–189.

Nunnally, J.C. & Bernstein, I. H. (1994). *Psychometric Theory.* NewYork: McGraw–Hill.

Olmez, I., & Cohen, A. (2018). A Mixture Partial Credit Analysis of Math Anxiety. *International Journal of Assessment Tools in Education*, *5*(4), 611-630. Retrieved from https://ijate.net/index.php/ijate/article/view/565

Peker, M. & Ertekin, E. (2011). The relationship between mathematics teaching anxiety and mathematics anxiety. *The New Educational Review*, *23* (1), 213–226.

Posamentier, A. S. & Stepelman, J. (1986). *Teaching secondary school mathematics.* Ohio: Charles E. Merrill Publishing Company.

Richardson, M. F. (1980). *An Assessment of Mathematics Anxiety Levels among Adult Basic and Adult Secondary Students* (Unpublished Doctoral Thesis). The University of Georgia, Athens. Retrieved from ProQuest Dissertations & Theses Global.

Ryang, D. (2012). Exploratory analysis of Korean elementary pre-service teachers' mathematics teaching efficacy beliefs. International Electronic Journal of Mathematics Education, 7(2), 45–61.

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

Schumacker, R. E. & Lomax, R. G. (2010). *A beginner's guide to Structural Equation Modeling* (3rd ed.). NJ: Lawrence Erlbaum Associates.

Skiba, A. (1990). Reviewing an old subject: Math anxiety. *Mathematics Teacher, 83*(3), 188–189.

Sloan, T. R. (2010). A Quantitative and qualitative study of math anxiety among pre-service teachers, *The Educational Forum*, 74(3), 242–256.

Sparks, S. D. (2011). Researchers Probe Causes of Math Anxiety. *Education Week*, 30(31). Retrieved from http://www.edweek.org

Stuart, V. (2000). Math curse or math anxiety? *Teaching Children Mathematics*, 6, 30-340.

Sümer, N. (2000). Yapısal eşitlik modelleri: temel kavramlar ve örnek uygulamalar. [Structural equation modelling: Basic concepts and best practices] *Türk Psikoloji Yazıları* [*Turkish Psychology Writings]*. 3(6), 49–74.

Swackhammer, L., Koellner, K., Basile, C. & Kimborough, D. (2009). Increasing the self-efficacy of inservice teachers through content knowledge. *Teacher Education Quarterly,* 36(2), 63-78.

Swars, S., Daane, C. & Giesen, J. (2006). Mathematics anxiety and mathematics teachers' efficacy: What is the relationship in elementary pre-service teachers? *School Science and Mathematics*, *106* (7), 306–315.

Şencan, H. (2005). *Sosyal ve Davranışsal Ölçümlerde Güvenlik ve Geçerlik.* Ankara: Seçkin Yayınları.

Tan, Ş. (2009). KR-20 ve Cronbach Alfa Katsayılarının Yanlış Kullanımları [Misuses of KR-20 and Cronbach's Alpha Reliability Coefficients] . *Eğitim ve Bilim [Education and Science], 34* (152), 101–112.

Tavşancıl, E. (2006). *Tutumların Ölçülmesi ve SPPS ile Veri Analizi* (3.Basım).Ankara: Nobel Akademik Yayıncılık.

Tezbaşaran, A. A. (1997). *Likert Tipi Ölçek Geliştirme Kılavuzu* (2.Basım).Ankara: Türk Psikologlar Derneği Yayınları.

Timm, N. H. (2002). *Applied Multivariate Analysis.* New York, NY: Springer.

Tobias, S. (1978). *Overcoming Math Anxiety.* Newyork: Norton.

Tobias, S. (1990). Math Anxiety: An Update. *NACADA Journal, 10*(1), 47–50.

Trilling, B. & Fadel, C. (2009). *21st Century Skills: Learning for Life in Our Times*. San Francisco, CA: John Wiley & Sons.

Uusimaki, L. & Nason, R. (2004). Causes underlying pre-service teachers' negative beliefs and anxieties about mathematics. *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education*, 4, 369–376.

Vinson, B. (2001). A comparison of pre-service teachers' mathematics anxiety before and after a methods class emphasizing manipulatives. *Early Childhood Education Journal*, *29* (2), 89–94.

Voogt, J. & Roblin, N. P. (2010). *21st Century Skills*. Enschede: University of Twente.

Vukovic, R. K., Kieffer, M. J., Bailey, S. P. & Harari, R. R. (2013). Mathematics anxiety in young children: Concurrent and longitudinal associations with mathematical performance. *Contemporary Educational Psychology, 38*(1), 1 - 10. http://dx.doi.org/10.1016/j.cedpsych.2012.09.001

Wu, S. S., Willcutt, E. G., Escovar, E. & Menon, V. (2014). Mathematics achievement and anxiety and their relation to internalizing and externalizing behaviors. *Journal of Learning Disabilities*, *47*(6), 503–514.

Zettle, R. D. & Houghton, L. L. (1998). The relationship between mathematics anxiety and social desirability as a function of gender. *College Student Journal*, *32*, 81-86.

Zettle, R. & Raines, S. (2000). The relationship of trait and text anxiety with mathematics anxiety. *College Student Journal*, *34* (2), 246.

APPENDIX

Mathematics Teaching Anxiety Scale (MTAS-Turkish version) for Prospective Primary SchoolTeachers

1	I'm worried that I cannot motivate the students due to my prejudices against mathematics. *Matematiğe yönelik önyargılarımdan dolayı öğrencileri motive edemeyeceğim endişesi yaşarım.*
2	I feel uncomfortable in mathematics since I do not have enough experience. *Yeterli deneyime sahip olmadığım için matematik dersinde kendimi huzursuz hissederim.*
3	I am afraid that the level differences of the students in mathematics may affect my teaching pace. *Matematik dersinde öğrencilerin düzey farklılıklarının ders işleme hızımı etkilemesinden korkarım.*
4	I am afraid that students with fewer interests in mathematics may reduce the interest of other students. *Matematik dersine ilgisi az olan öğrencilerin diğer öğrencilerin ilgisini azaltmasından korkarım.*
5	The thought that the student cannot comprehend when I turn a concept into a mathematical sentence (e.g. 2 + 3) makes me anxious. *Bir kavramı matematiksel cümleye (ör: 2+3) dönüştürdüğümde öğrencinin anlayamayacağı düşüncesi beni tedirgin eder.*
6	I'm worried about not using the appropriate method and technique in mathematics. *Matematik dersine uygun yöntem ve tekniği kullanamama endişesi yaşarım.*
7	A rise in the level differences among my students while teaching mathematics worries me. *Matematik dersini işlerken öğrencilerim arasında düzey farklılıklarının artmasından endişelenirim.*
8	I am anxious since I believe that I do not have sufficient knowledge about teaching mathematics. *Matematik öğretimine yönelik yeterli bilgiye sahip olmadığımı düşündüğümden endişelenirim.*
9	When a student does not understand mathematical operations, I get anxious about how to explain them. *Matematiksel işlemleri öğrenci anlamadığında nasıl açıklayacağım endişesi yaşarım.*
10	I feel insecure about the thought that my students having level differences in mathematics can isolate themselves from the class eventually. *Matematik dersinde düzey farklılıkları olan öğrencilerimin zamanla kendilerini sınıftan soyutlayabilecekleri düşüncesi beni huzursuz eder.*
11	I feel anxious that I cannot finish the outcomes of the mathematics curriculum on time. *Matematik programındaki kazanımları zamanında bitiremeyeceğim endişesi yaşarım.*
12	I am afraid that school administrators will criticize me if I cannot catch up with the mathematics curriculum. *Matematik programını yetiştiremezsem okul yöneticilerinin beni eleştirmesinden korkarım*
13	I get anxious about designing activities that are appropriate for my students’ level in mathematics. *Matematik dersinde öğrencilerimin düzeyine uygun etkinlik hazırlama endişesi yaşarım.*
14	I feel anxious while considering students’ individual differences in teaching mathematics. *Matematik öğretirken bireysel farklılıkları göz önünde bulundurma zorunluluğu beni endişelendirir.*
15	I am afraid that families will criticize me if I cannot catch up with the mathematics curriculum. *Matematik programını yetiştiremezsem ailelerin beni eleştirmesinden huzursuz olurum.*
16	I'm worried about not enabling my students' to engage in mathematics actively. *Öğrencilerimin matematik dersine aktif katılımını sağlayamama endişesi yaşarım.*
17	I feel worry that I do not know how to teach mathematical concepts to students. *Matematik kavramlarını kazandırırken nasıl öğreteceğimi bilmediğim için tedirgin olurum.*
18	I feel fear of being humiliated by the students if I cannot solve problems in mathematics. *Matematik dersinde problemleri çözemezsem öğrencilerin gözünde küçük düşmekten korkarım.*
19	The thought that the level differences of the students in mathematics may reduce the interest of attending the lesson disturbs me. *Matematik dersinde öğrencilerin düzey farklılıklarının derse olan ilgiyi azaltacağı düşüncesi beni rahatsız eder.*

Chronbach’s Alpha Coefficient: 0.93

International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 522–538
https://dx.doi.org/10.21449/ijate.626053

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

Development of a "Perceived Stress Scale" Based on Classical Test Theory and Graded Response Model

MetinYaşar [1,*]

[1] Pamukkale University, Faculty of Education, Kınıklı Campus, 20070, Denizli, Turkey

ARTICLE HISTORY

Received: 28 May 2019

Revised: 08 August 2019

Accepted: 25 September 2019

KEYWORDS

Classical test theory,
Item response theory,
Item difficulty,
Item discrimination index,
Item invariance

Abstract: The main purpose of this study is to develop a perceived stress scale based on Classical Test Theory (CTT) and Graded Response Model (GRM); to compare the parameters of the items in the scale that are tried to be developed according to both models, and to determine under which theory the measurement tool produces more reliable and valid results according to these compared item parameters. The item discrimination parameter value calculated according to CTT ranges from 0.472 to 0.735. On the other hand, item discrimination parameter values calculated under GRM vary between 1.062 (Item 15) and 2.606 (Item 2). Correlations between item thresholds were tested and the calculated correlation coefficients were; r =0.840 for β1 ($p<0.01$), r = 0.947 for β2 ($p<0.01$), r = 0.713 for β3 ($p<0.05$), and r = 0.559 for β4 ($p<0.05$) respectively. It can be assumed that these values not only support the item invariance of the items in the scale, but also show that the GRM is suitable for the data used for the scaling of the items. The reliability coefficient of the scale, in terms of internal consistency, was calculated as 0.919 according to the CTT, while the marginal reliability coefficient calculated as 0.931 according to GRM. Both reliability coefficients are quite high. In conclusion, there is a high correlation between the item parameters calculated according to both approaches, and the perceived stress scale (PSS) that is being developed can measure the desired features.

1. INTRODUCTION

The measurement and evaluation carried out in the education system are used in planning the education, improving the quality of the education system, organizing the content used in education, and activating the mechanism necessary for reviewing the content that is not related to the determined objectives. In addition, it serves to determine the adequacy of the individuals to be measured according to the determined objectives, to compare the performance or academic achievement of the students depending on the purpose, and to provide the necessary inputs for the training of individuals in line with the determined goals.

Measurement in the broadest sense, is defined as the process of observing any quality of individuals and expressing the results of observations by numbers or symbols (Turgut, 1992; Turgut & Baykul, 1992). Measured qualifications of individuals may be cognitive, effective, or psychomotor; such as an individual's academic achievement in any subject, attitudes towards anything, or psychomotor skills. At this point, when the relevant characteristics of individuals

CONTACT: Metin YAŞAR ✉ myasar@pau.edu.tr Pamukkale University, Faculty of Education, Kınıklı Campus, 20070, Denizli, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

are to be measured, the effectiveness of the measurement and evaluation becomes important. The main aim of the researchers in the field is to contribute to the development of effective and new approaches to increase the effectiveness of measurement and evaluation, and to enable the development of measurement tools that will reveal the values closest to the actual magnitudes of the features to be measured.

Two important theories are used intensively in order to develop the measurement tools used to determine the cognitive, effective and psychomotor characteristics of individuals. One of these theories is known as Classical Test Theory (CTT) and the other is known as Item Response Theory (IRT).

1.1. Classical Test Theory (CTT)

Classical Test Theory is a simple theory that explains the observed score of the test with the actual score and the measurement error. Despite the weak assumptions of classical test theory that can be met by data sets from many applications, it is used in a wide range of applications that require test development and interpretation of test scores (Hambleton & Swaminathan, 1989). Until the statistical approach of Lord and Novick (1968), later known as Item Response Theory (IRT), which describes latent properties test scores, CTT continued being the predominant (Sijtsma & Junker, 2006; Seungho-Yang, 2007) theory of explanation and interpretation of test scores (Köse, 2015). Based on the test results and the measurement results obtained from the application, CTT was preferred more due to the ease of estimating the parameters of the item and the small number of assumptions (Kelecioğlu, 2001, cited in Kan, 2006). Although Classical Test Theory is based on Spearman's (1905) basic equation, it accepts the existence of both the actual score and the error score of the observed property of the individual.

The basic equation of classical test theory is expressed as follows:

$$X = T + E \qquad (1)$$

X = Observed Score

T: True Score

E: Random Error

According to the assumption of Classical Test Theory, the characteristics of an individual are fixed, and the variation in observed scores results from random errors, which are the result of various factors such as failure or chance of success (Doğan & Tezbaşaran, 2003).

Furthermore, according to the CTT, the item difficulty index $(\boldsymbol{p})$ and item discrimination index $(\boldsymbol{r_{jx}})$ are used as the starting point for an ideal test (high reliability and validity). It is possible to estimate test statistics based on item statistics. In Classical Test Theory, the scores of individuals vary according to the difficulty level of the test items, and thereby to the test as a whole. However, the calculation of a standard error score can be considered as one of the weaknesses of this theory, as if the error score of the individuals involved in the test scores obtained from a test is the same for the whole group.

Because of the easy-to-meet assumptions of the CTT, it has been easily used in the past to solve many measurement problems in test development. Nowadays, there are many tests of success, talent, personality etc. developed according to this theory. Although Classical Test Theory is used frequently nowadays, it has some weak assumptions. Therefore, there are many criticisms about the development, implementation and evaluation of tests used in education and psychology based on this theory. One of these criticisms is that the frequently used item statistics depend on the selected sample and are influenced by the sample (Lord & Novick, 1968; Lord, 1980; Hambleton & Swaminathan, 1985; Crocker & Algina, 1986; Gelbal, 1994; Embretson & Reise, 2000; Nartgün, 2002; Doğan & Tezbaşaran, 2003; Köse, 2015). The fact that CTT has weak assumptions can also be seen as an advantageous feature of the theory over IRT (Hambleton & Jones, 1993). An example of the advantageous features of CTT may be the fact that IRT applications require large samples, while CTT applications can be performed without requiring very large samples (Bichi, Embong, Mamat & Maiwada, 2015).

CTT does not include latent variables: operationally, although the actual score is not empirically observable, it can be defined as the average score in the infinite equivalent number of repetitions (Lord & Novick, 1968). Lord (1953) stated that observed scores and true scores are not synonymous with ability scores of individuals, whereas skill scores are more basic and independent of the test or test items within the test, but observed scores and actual scores are dependent on the test (Hambleton &Jones, 1993: cited in Sünbül & Erkuş, 2013).

1.2. Item Response Theory (IRT)

Based on the limitations of CTT, it is known that in the late 1930s, properties of the theory known as item reaction theory began to be discussed in order to eliminate the disadvantages of these limitations, and in 1940 Tucker was the first to use the concept of item characteristic curve, which was accepted as one of the most important features of Item Response Theory (Doğan & Tezbaşaran, 2003).

Item properties in latent-trait model, depending on the selected model, are: (1) parameter $\boldsymbol{b}$, the ability level best measured by the item; or in addition to previous one, (2) parameter $\boldsymbol{a}$, which provides information about the quality of the item; or in addition to the previous two, (3) parameter $\boldsymbol{c}$, the likelihood of the item being answered correctly by chance. Parameter $\boldsymbol{b}$ specified in the first item of the list is the parameter of the Rasch dichotomous model, and the One-parameter Logistic Model; the parameters specified in the second item are parameters of the two-parameter logistic model; In the third item, the parameters specified in the third item are parameters of the three-parameter logistic model (Gelbal, 1994). One of the differences between item statistics in CTT and item parameters in IRT is that, $\boldsymbol{p_j}$ and $\boldsymbol{r_{jx}}$ are obtained from the group in which the test is developed in CTT, whereas $\boldsymbol{b}$ and $\boldsymbol{a}$ parameters in IRT are obtained from a mathematical distribution function according to the selected model. According to many authors, the superiority of IRT over CTT is that item properties can be calculated independently from the group by means of this function (Lord & Novick 1968, Hamblethon & Swaminathan 1985).

Besides IRT's aforementioned advantage over CTT, there are similarities between these two theories. Item difficulty index ($\boldsymbol{p_j}$) in CTT and parameter ($\boldsymbol{b}$) which is the ability level best measured by the item in latent-property theory, and the item discriminatory power index ($\boldsymbol{r_{jx}}$) in CTT and parameter a which provides information about the quality of the item have the same meaning reciprocatively. Equations for the transition from each of these two parameter pairs to the others are given by Lord and Novick (1968). These equalities express the similarities between IRT and CTT. Weiss (1983) touches on this similarity in another aspect and states that IRT is in fact derived from CTT, and that CTT is a very simple form of IRT (Gelbal, 1994).

1.3. Graded Response Model (GRM)

The Graded Response Model (GRM) is generally known as a model used in the analysis of personal data (Embretson & Reise, 2000; Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Robie, Zickae & Schmit, 2001; La Huis & Copeland, 2009). GRM is the most commonly applied item response model to intermittent scale data (Lautenschlager, Meade & Kim 2006). In GRM, there are $\boldsymbol{m}$ ranked categories specific to each item. Items that can be scored as multiple are considered as categorical items similar to items that can be scored as binary (Köse, 2015; Bilgen & Doğan, 2017), and they have more than two response categories. Values separating these categories are expressed as limit values or threshold values. Instead of calculating one item difficulty parameter for each item under GRM, the category response threshold value for *m-1* item categories is calculated. If the scale items are composed of 5-point Likert type items, 4 threshold values or limit values for each item are calculated. These limit values are sorted in an ascending order. Under GRM, each item is represented by two item parameters. The first of these parameters is called item discrimination parameter and the second is called item difficulty parameter. The item discrimination parameter, as a function of the latent-property to be measured, can also be considered as the power or probability of changing the response of in the categories (In practice, a high discrimination parameter value means that the probability of a correct response increases more rapidly as the ability or latent trait increases).

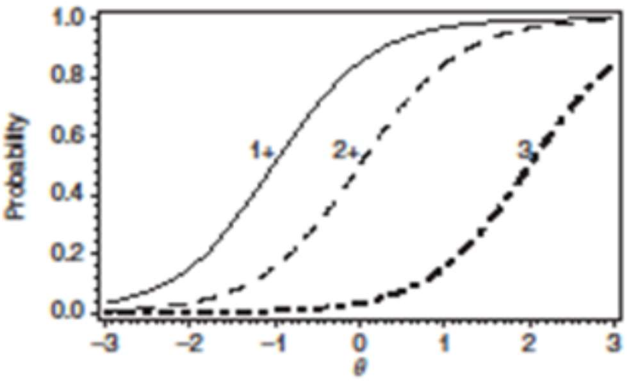


Figure 1. *GRM model for a 4-category item (ranked between 0-3). (Excerpt from DeMars, 2010).*

As can be seen in the item given in Fig. 1, similar functions, such as an item characteristic curve, can be drawn for each category. de Ayala (1993) used the name Process Characteristic Curves (PCC) for the curves in Figure 1 (cited in, DeMars, 2010), while Embretson and Reise (2000) used the name Process Characteristic Curves (PCC). In GRM, each item is defined by two parameters. The first is the item difficulty level and the second is the item discrimination index.

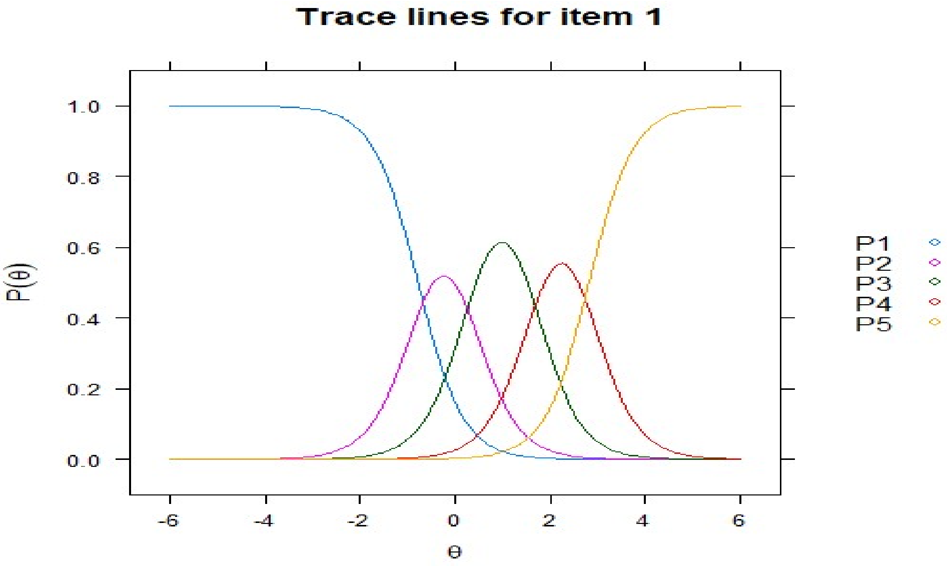


Figure 2. *All category/score possibilities for a 5-category item. These probabilities are calculated using item limit values or item threshold values.*

Sijtsma and Meijer (2007) calls the curves shown in Figure 2, the category response function (CRF), Muraki (1992) calls them the item category response function (ICRF) and Ayala and

Sava-Bolesta (1999) calls them the option response function (ORF) (cited in, DeMars, 2010). Although the mathematical function of GRM is very similar to the 2PL function, it cannot be calculated directly from the 2PL model. This is because only one $\boldsymbol{b}$ parameter is calculated in the 2PL function. GRM's only difference from the 2PL function is that it has multiple $\boldsymbol{b}$ parameters. For a ranked Likert-type item under GRM, a parameter $\boldsymbol{b}$ is calculated for each of the remaining categories except the first category.

$$\boldsymbol{P}_{ik}^{*}(\boldsymbol{\theta}) = \frac{\boldsymbol{e}^{1.7\boldsymbol{a}_i(\boldsymbol{\theta}-\boldsymbol{b}_{ik})}}{1+\boldsymbol{e}^{1.7\boldsymbol{a}_i(\boldsymbol{\theta}-\boldsymbol{b}_{ik})}} \qquad (2)$$

In Equation 1, $P_{ik}^{*}(\theta)$ indicates the probability of the i item scoring at or above the k category (in a specified θ and item parameters), $\boldsymbol{a}_{\mathrm{i}}$ indicates discrimination parameter for i item and $\boldsymbol{b}_{ik}$ indicates difficulty parameter of i item in k category. When an item's $b_{i1} = -1.0$ 50% of the individuals with $\theta = -1.0$ will score 1 or higher. In the above equation, letter i, which is shown as a subscript, indicates item i, and * or + signs added to the P expression indicate the possibility of receiving/selecting points in or above that category, not the probability of making/choosing points (DeMars, 2010). The a-parameter in the above equation can be interpreted in a similar sense to the a-parameters in the two-category items. Although the a-parameter in the GRM is widely used as the item discrimination parameter, some researchers do not prefer to use it in multi-category items (Embretson & Reise, 2000). DeMars (2010), on the other hand, considers the degrees of items that differentiate individuals having different θ values in multi-category items, as a function of relative locations of a-parameter and b-parameters. He also emphasizes that it is very common to use the 'item separator parameter' expression for the a-parameter.

Although Item Response Theory is accepted as a powerful test theory according to classical test theory and it is very popular, model data alignment needs to be ensured. Although there is no definite test for model-data fit in IRT, the number of iterations and parameter invariance of items can be considered as methods that can provide information about item-data fit in the item analyses (Rubio et al., 2009, cited in, Köse, 2015).

Since IRT is a theory based on each item that constitutes the test, each item in the test is assumed to measure a latent property. As a result, the amount of information for a single item can be calculated at any skill level and indicated by I$_{\mathrm{i}}$ *(θ)*. Therefore, the level at which an item can make the most sensitive measurement can be considered as the place where the item corresponds to the level of difficulty.

It can be said that stress is one of the most frequently complained subjects in today's society. While stress affects people in such a negative sense, there is no common definition of stress in studies. Many definitions are made for understanding stress and efforts are made to explain it with anthropological, physiological, endocrinological, sociological and psychological approaches. On the other hand, it is reported that the existence of different explanations and approaches creates a confusion and makes it difficult to understand the connections between these approaches (Tatar, Saltukoğlu & Özmen, 2018).

> Approaches or conceptualization efforts to explain stress are classified according to different criteria. One of these classifications is grouped under three titles: Response, Stimuli, and Transactional. The Response focuses on physical processes; the Stimuli focuses on environmental stimuli or external demands; the Transactional focuses on cognitive processes. Another classification is divided into two categories as Biological and Psychosocial. The Biological approach includes the physiology and endocrinology-based response approach, and the psychosocial approach includes stimulant and process approaches. The biopsychosocial model (BPS) is presented as an approach that combines these two approaches in a single framework. (Tatar, Saltukoğlu & Özmen, 2018).

Today, it is a known fact that educators, especially teachers as an indispensable part of education, experience a very high level of stress. This study aims to develop a scale that can determine the level of stress levels the teachers experience in the education system while performing their professional duties. The purpose of the developed measurement tool is to have the characteristics that can be used to determine the perceived stress level of teachers. CTT and GRM assumptions, which are briefly explained above, were used in the development of PSS.

2. METHOD

2.1. Participants

In order to develop the Perceived Stress Scale (PSS), a draft scale consisting of 51 items was applied to 475 volunteering teachers working at different levels in schools affiliated with the Ministry of National Education in Denizli, Turkey.

2.2. Data Collection Tool

In this study, there is an effort to develop a new scale in order to reveal the perceived stress levels of teachers by using the CTT and GRM approaches instead of working with any existing scale. The scale was developed as a 5-point Likert-type scale, and the literature was reviewed before writing the items in the scale. Reviewed studies include: The Adaptation of the Perceived Stress Scale into Turkish: A reliability and Validity Analysis (Eskin, Harlak, Demirkıran & Dereboy, 2013), The Effect of Perceived Organizational Support and Work Stress on Organizational Identification and Job Performance (Turunç & Çelik, 2010), Framing Focus of Control & Workaholism Positively With Reference to Perceived Stress (Akdağ & Yüksel, 2010), The Relationship Between the Perceived Stress Level and the Stress Coping Strategies in University Students (Savcı & Aysan, 2014), Turkish Adaptation of Perceived Stress Scale, Bio-psycho-social Response, and Coping Behaviours of Stress Scales for Nursing Students (Karaca et al., 2013), Reliability and Validity of the Turkish Version of Perceived Stress Scale (Erci, 2006), Analysing the Perceived Stress Level of Teachers with Regards to Some Variables (Şanlı, 2017), The Sources of Stress, Coping, and Psychological Well-Being among Turkic and Relative Societies' Students in Turkey (Otrar, Ekşi, Dilmaç, & Şirin, 2002). Based on these studies, 51 items were written for the Perceived Stress Scale (PSS).

Fifty-one items in the perceived stress scale were ranked from the most negative expression 'strongly disagree (1)' to the most positive expression 'strongly agree (5)'. Before applying the 51-item Perceived Stress Scale (PSS) to the study group, the teachers were informed about the purpose of the scale to be applied to them. Furthermore, a motivating explanation was given to the study group, informing them that their personal information won't be required, in order to encourage them to select the most appropriate option by reading the items in a more sensitive way. The 51-item PSS draft was applied to 475 teachers, and as 26 teachers in the study group left many items unanswered, their answers are not included in the study. The feature that differentiates the Perceived Stress Scale (PSS) that was developed in this study from similar scales is that there is no scale developed based on both CTT and GRM in the literature.

2.3. Data Analysis

The data obtained from the application of Perceived Stress Scale (PSS) draft were first entered into SPSS 22.0 environment in order to perform the necessary analyses according to CTT. The data obtained from the study group were analyzed using SPSS 22.0 and R programs, according to CTT and GRM respectively. Item discrimination index and item difficulty index were calculated as item parameters according to CTT. While item-total correlations were used as item discrimination parameter, item averages were taken into account for item difficulty parameter. Furthermore, the Cronbach alpha coefficient was calculated for reliability in terms of internal consistency of the scale that trying to be developed according to CTT. For GRM, firstly, the graded response model developed by Semejima (1969) was used. Within the scope of the analysis of the raw data obtained as a result of the application of the PSS; the items with item-total correlation values below 0.40 or were overlapping (according to CTT), and items that violate local independence were (according to the IRT) excluded from the scale. After the unsuitable items were removed from the scale according to both theories, a final scale of 16 items emerged. Statistical analyzes of perceived stress scale (PSS) are explained in more detail in the Results section.

3. FINDINGS

The aim of this research is to develop a scale that is highly reliable and valid for both the CTT and the IRM under the IRT, which can determine the degree of perceived stress levels of the teachers working in the education system. In this context, firstly the item discrimination and item difficulty levels were calculated as item statistics, based on the measurement results obtained from the answers given by the respondents in the study group according to CTT. In such scales, it is useful to consider that the item difficulty level is different from the difficulty level of an item in an achievement test. The item difficulty level here should be seen as the difficulty of decision-making in the preference of expressions in ranked categories. The difficulty $(\boldsymbol{p_j})$ level of any item in the achievement test is known as the correct response rate of that item. However, there is no ratio of correct answers in ranked Likert-type scale items. It would be useful to consider the difficulty here as the difficulty the participant has in choosing the item that describes the situation best. Item-total correlations of scale items were calculated as item discrimination parameter. The high item-total correlations of the items in the scale ensure that the measurements are close to the actual value. Cronbach's alpha coefficient was also calculated to determine the internal consistency of the items in the scale. Cronbach's alpha reliability coefficient was calculated as 0.919 and it is a quite high value. The values of the item parameters calculated according to CTT are as in Table 1.

Many studies in the field claim that IRT has superior features compared to CTT (Lord, 1980; Hambleton & Swaminathan, 1985; Blood, 2006; Gelbal, 1994; Doğan & Tezbaşaran, 2003; Nartgün, 2002). Although it is claimed that IRT has many positive advantages over CTT, it is stated that the power of IRT is based on one-dimensionality and depends on meeting this assumption (Lord, 1980; Hambleton & Swaminathan, 1985; Kan, 2006). It is claimed that, as an evidence for its one-dimensionality, the scale should have a dominant factor (Lord, 1980; Hambleton & Swaminathan, 1985; Kan, 2006; Doğan & Tezbaşaran, 2003; Nartgün, 2002; Bichi & Talib, 2018). The eigenvalue graph, which is one of the methods used to determine the one-dimensionality of the scale, is one of the most effective methods in revealing the dominant factor (Kan, 2006; Köse, 2015). In addition, as a measure of the one-dimensionality of the scale, the scale is assumed to be one-dimensional if there is at least two-times difference between the size of the eigenvalue of the first component and the the second component (Gelbal, 1994). If the first dominant factor explains 20% or more of the variance, the scale is assumed to be one-dimensional (Lee, 1995; cited in Köse, 2015).

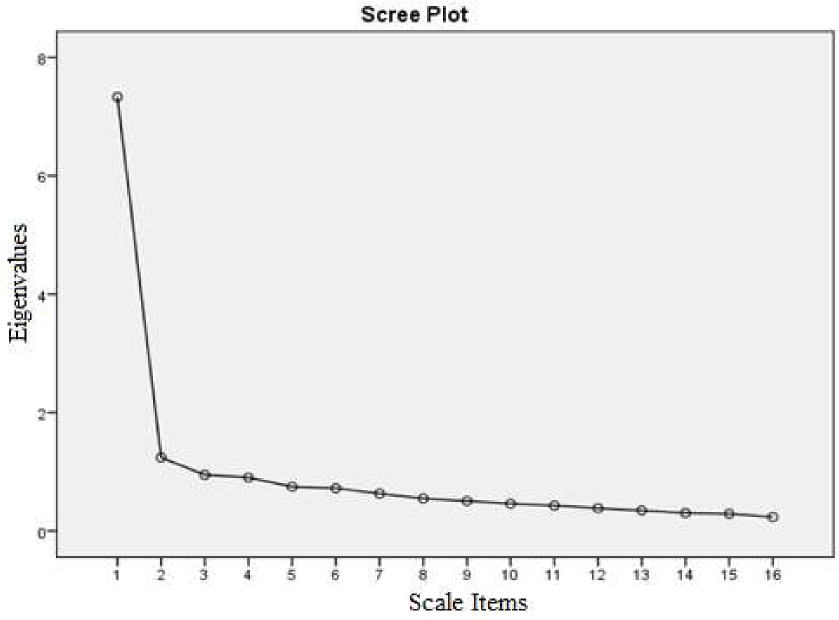


Figure 3. *Eigenvalue graphic*

In Figure 3, the eigenvalue graph of the scale data shows the factor structure of the scale. In order to say that the scale is one-dimensional, there must be at least twice the difference between the first factor and the second factor. This situation has also been realized on the scale that is being developed in the study. There is almost six times the difference between the eigenvalues of the first and the second factor. Another criterion is that the first factor explains at least 20% or more of the variance; in this study, the first factor explains 45.83% of the variance. Therefore, it can be said that the Perceived Stress Scale (PSS), which is tried to be developed according to CTT, is a one-dimensional scale with high reliability.

The second theory used in the development of Perceived Stress Scale (PSS) is IRT. According to the assumptions of GRM under IRT, the data obtained from the study group were analyzed using the R program. First, item discrimination parameter ($\boldsymbol{a_i}$) and then four item threshold values (difficulty parameter) were calculated. The high level of item discrimination parameters indicates that individuals can be better distinguished from each other according to their ability levels. It is therefore expected that the discriminant parameters of the scale items would be as high as possible. On the other hand, the items with low $\boldsymbol{a_i}$ parameter values are insufficient to distinguish individuals according to their ability levels in terms of measured characteristic. The high $\boldsymbol{a_i}$ values of the items in the scale contribute positively to the item information function and thus to the test information function. Table 1 shows item and test parameters obtained according to both CTT and GRM. The marginal reliability coefficient calculated under ATM is calculated as .931 and the curve of this marginal reliability coefficient is given in Figure 4.

Table 1. *The parameters predicted under CTT and GRM*

	CTT $\boldsymbol{\alpha}$ = .919		GRM $\boldsymbol{\alpha}$ = .931									
Item	α_{CTT}	$\mathrm{b}_{\mathrm{CTT}}$	α	S_E	β_1	S_E	β_2	S_E	β_3	S_E	β_4	S_E
M1 (2)	.667	2.2472	2.123	.176	-0.780	.089	0.303	.075	1.652	.129	2.828	.247
M2 (3)	.735	2.3519	2.606	.212	-0.947	.087	0.208	.069	1.466	.108	2.475	.196
M3 (4)	.584	2.2606	1.675	.147	-0.988	.109	0.476	.087	1.655	.145	3.433	.361
M4 (5)	.690	2.5523	2.098	.171	-1.467	.118	0.054	.074	1.259	.106	2.359	.192
M5(6)	.515	2.8864	1.266	.120	-2.047	.198	-0.491	.104	0.831	.116	2.471	.244
M6 (7)	.639	2.3964	1.849	.155	-1.104	.108	0.174	.078	1.490	.127	2.843	.254
M7(8)	.710	2.5323	2.253	.180	-1.146	.100	0.007	.072	1.137	.097	2.408	.196
M8(10)	.650	1.9621	2.079	.177	-0.440	.081	0.88	.090	2.051	.162	2.726	.242
M9(11)	.720	2.2784	2.408	.194	-0.708	.083	0.332	.071	1.335	.105	2.475	.200
M10(12)	.599	2.0290	1.688	.152	-0.592	.093	0.778	.096	2.105	.183	3.285	.338
M11(13)	.635	2.2027	1.791	.156	-0.812	.097	0.554	.086	1.740	.146	2.523	.221
M12(28)	.641	2.4922	1.758	.152	-1.283	.120	0.100	.080	1.458	.128	2.317	.200
M13 (29)	.596	2.6370	1.534	.138	-1.598	.150	-0.147	.086	1.313	.130	2.463	.225
M14(32)	.543	2.6036	1.296	.123	-1.597	.163	-0.088	.095	1.498	.152	2.728	.267
M15 (35)	.472	2.4655	1.062	.115	-1.627	.190	0.003	.107	2.050	.232	3.596	.414
M16(45)	.481	2.5056	1.259	.124	-1.487	.158	0.095	.096	1.554	.162	3.158	.330

In Table 1, item-total correlation in the factor analysis results performed under CTT is considered as item discrimination parameter. Here, the item discrimination parameter value calculated according to CTT ranges from 0.472 to 0.735. On the other hand, item discrimination parameter values calculated under GRM vary between 1.062 (Item 15) and 2.606 (Item 2). The item discrimination parameters calculated under GRM for the items in the scale are quite high. Correlation between item discrimination parameters (that are calculated according to CTT and GRM) was tested to determine whether there was a significant relationship. The test showed a relationship (r = 0.970) between CTT and GRM item discrimination parameters (p <0.01). The scale items clustered under a dominant dimension may be the cause of the high item separation parameters obtained under both approaches (Köse, 2015).

As shown in Table 1, item difficulty levels and item threshold values were examined under GRM and as expected, threshold parameters of the items were ranked from the lowest to the highest value. In the table, β_1 shows the lowest and β_4 the highest threshold parameter for each item. Threshold parameters of the 1st Item in the scale were calculated as $\beta_1 = -0.780$ and $\beta_4 = 2.828$. According to these parameter values, the ability level to correctly answer this item in Category 1 with a 50% probability is θ = -0.780, while the ability level to respond with a 50% probability in Category 5 is θ = 2.828.

The most important advantage of latent traits theory is the invariance of item parameters. Since sufficient evidence is not provided for item invariance in studies (Fan 1998; Hambelton et al. 1991; Somer 1998; Stage 1998; Nartgün 2002), it remains a controversial issue (Doğan & Tezbaşaran, 2003). Since the determination of the invariance property of the item parameters is seen as an important requirement according to IRT, in this study, in order to test the invariance of the item parameters, the study group was randomly divided into two groups by means of SPSS-DATA-SELECT CASE and the evidence for the invariance of the items in the scale was obtained from the level of the relationship between the item parameters obtained from the two semi-groups. Since the study group in this study was divided into two, the item discrimination parameter of the measurement results obtained from both groups and the threshold values of each item in the test were calculated. The correlation between item discrimination parameters was calculated as r = 0.737 according to the results of two half-groups (p <0.05 Correlations between item thresholds were tested and the calculated correlation coefficients were; r =0.840 for β1 (p<0.01), r = 0.947 for β2 (p<0.01), r = 0.713 for β3 (p<0.05), and r = 0.559 for β4 (p<0.05) respectively. These values support the item invariance of the items in the scale, and also show that the GRM is suitable for the data used for scaling the items.

Local independence, which is one of the important assumptions of IRT, means that individuals' responses to items are statistically independent and unrelated when the ability to influence test performance is kept constant (Reckase, 2009; Erkuş, Ö. Sünbül, Sünbül, Yormaz & Dereboy, 2017; Bilgen & Doğan, 2017). In other words, local independence means that the responses to one item are independent of other items at a certain level of ability. Accordingly, local independence does not mean that there is no correlation between the items for all groups; however, it means that the responses to the item are independent at different skill levels. According to Lord and Novick (1968), it may be wrong to think that a group of test items would be independent according to the local independence approach. When differences between individuals' abilities are observed, there may also be positive relationships between test items. These relationships should not affect test scores at a fixed ability level. In order to meet the assumption of local independence, it is a necessity to meet the one-dimensional assumption. If the test has a one-dimensional property, it can also be assumed that it also meets the local independence assumption. If the responses to items in a one-dimensional model are not locally independent of each other, it causes another dimension dependency. Items that do not meet the assumption of local independence become overlapped items, and therefore give less information than the information it should provide. The tests used for local independence in studies usually focus on dependence between substance pairs. This dependence may not appear as separate dimensions unless it affects a large proportion of the items. This may not be determined by whether the test is one-dimensional. Although it is considered sufficient for a measurement tool to be one-dimensional to meet the assumption of local independence, some other methods are used to test local independence. One of these methods is the Q_3 test proposed by Yen (1984) in order to check the local independence between the pairs of items in the measurement tool. According to the Q_3 test, local independence is the calculation of the residues of the responses to each item for each individual based on the item parameter estimation. The residues mentioned here are the difference between the predicted and observed item parameters. After obtaining the residues, the linear correlation between the residues of items Q_3, i and j is calculated. Items that violate the assumption of local independence are found by examining the highly correlated items based on the correlation matrix obtained. Yen's (1984) recommendation to researchers is that if the linear correlation coefficient between the criteria items is greater than 0.20, they should approach these items as if they were violating local independence. In this study, using the R program, it was tested whether the items in the scale meet the assumption of local independence for the data obtained from the study group. As Yen (1984) suggested, Q_3 test was performed and according to the test results, items with a correlation value greater than 0.20 were excluded from the scale and local independence assumption was made for the items.

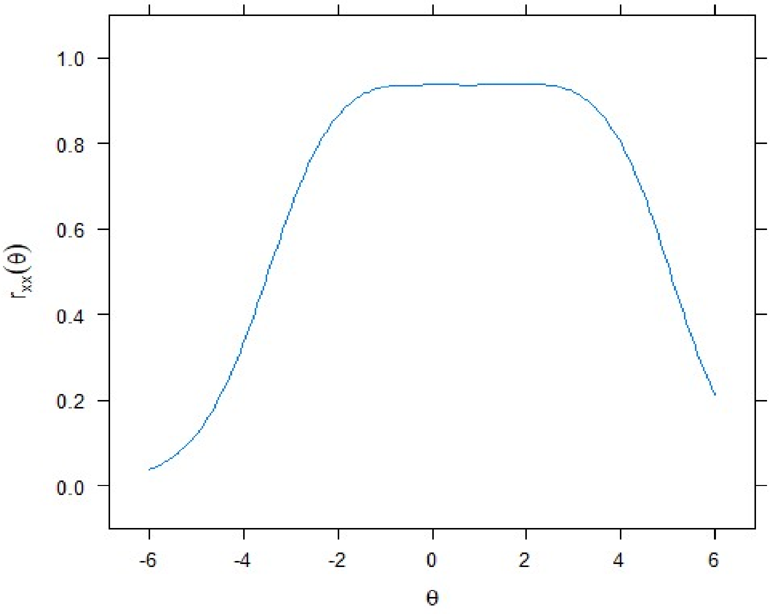


Figure 4. *Marginal reliability coefficient of PSS according to GRM*

One of the biggest criticisms of CTT is that a single coefficient of reliability is estimated and used for the entire range of capabilities tested. On the other hand, the information functions in IRT are used in the same sense even if they are not the exact equivalent of the reliability in CTT. Item information functions of 16 items in the scale were calculated. Item information functions are shown in Figure 5. When item information functions are examined, it can be seen that all items in the scale contribute to test information function at high level.

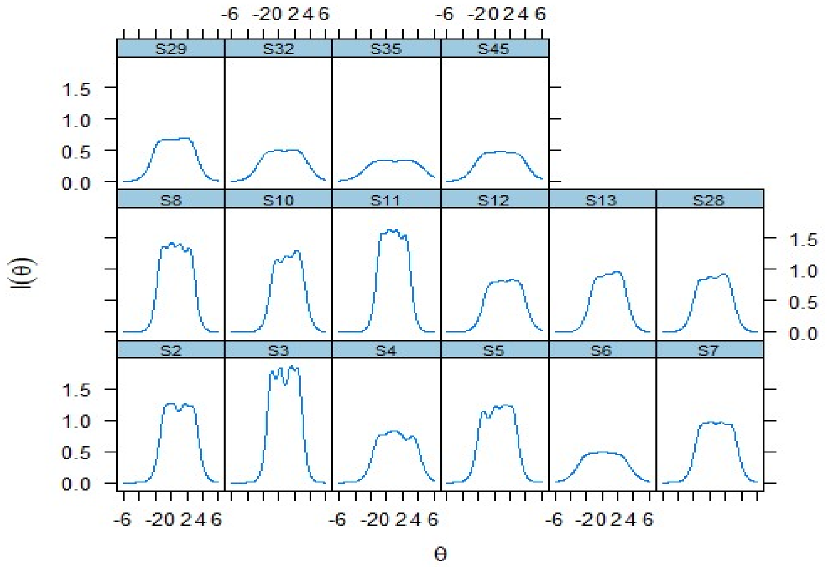


Figure 5. *PSS Item information functions*

Each scale item's contribution to the test information function was taken into account while calculating the test information function. The test information function is shown in Figure 6.

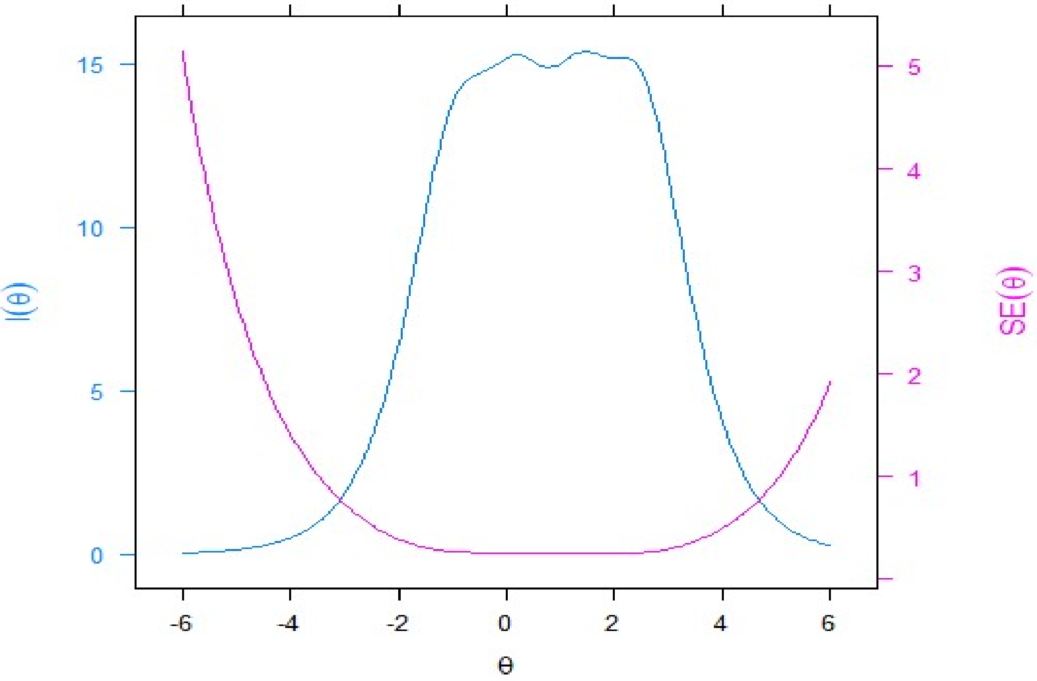


Figure 6. *Test information function*

The sum of the test items' information functions gives us the test information function. The test information function corresponds to about -1 and 2.4 skill levels according to GRM.

4. DISCUSSION and CONCLUSION

The main aim of this study is to develop a Perceived Stress Scale (PSS) for teachers using GRM, under both CTT and IRT. In the development of the perceived stress scale, item parameters were compared using both CTT and IRT. As Köse (2105) states, in order to make such comparisons, the obtained data must meet the one-dimensional assumption. For this purpose, the data obtained from the measurement process was subjected to exploratory factor analysis. Upon the examination of the findings obtained as a result of exploratory factor analysis, a significant difference was found between the eigenvalues of the first and the second factor. In addition to the large difference between the eigenvalues of the two factors, the first factor explains 45.83% of the variance in the study. If the first dominant factor explains 20% or more of the variance, the scale is assumed to be one-dimensional (Lee, 1995; cited in Köse, 2015). Therefore, it shows that the Perceived Stress Scale (PSS) that is being developed according to CTT is one-dimensional.

In order to determine whether there is a significant relationship between item discrimination parameters calculated under both CTT and GRM, correlation was calculated between both discrimination parameters. The test showed a relationship (r=0.970) between CTT and GRM item discrimination parameters (p <0.01). It can be said that there is a very high level of relationship between item discrimination parameters calculated according to both methods. It is an indicator that the same items should be on the scale according to both CTT and GRM. The findings obtained in this study are supported by Köse (2015) and Koch (1983). There is a parallel between item discrimination index values and item information functions. Items with high item discrimination index (Item 3, Item 11) have higher item information functions than others. On the other hand, among the 16 items in the scale, it is seen that the information function of item-35, which has the lowest item discrimination index value, is smaller than the information functions of other items.

The reliability coefficient of the scale, in terms of internal consistency, was calculated as 0.919 according to the CTT, while the marginal reliability coefficient calculated as 0.931 according to GRM. These reliability coefficients are quite high, and close to each other. Köse's findings (2015) support the findings of this study. In Köse's study (2015), values of 0.93 and 0.94 were obtained for CTT and GRM respectively. In this study, the results of the item parameters and reliability coefficients of the scale were found to be very similar to each other. Although the findings obtained from both approaches are similar, it can still be considered that GRM is one step ahead of CTT in its scale development effort. Because, in the analysis under GRM, test and item information functions make a great contribution to the researchers visually. This feature can be seen as an advantage.

As a result, perceived stress scale (PSS) has reliability and validity as a result of analyzes performed under GRM both in CTT and IRT framework. With the help of this scale, reliable and valid measurements of the perceived stress level of the participants can be made. This scale can be used to determine the perceived stress level of not only teachers, but also individuals working in other fields or university students.

ORCID

Metin YAŞAR https://orcid.org/0000-0002-7854-1494

5. REFERENCES

Akdağ, F., & Yüksel, M. (2010). İnsan kaynakları yönetimi açısından işkoliklik ve algılanan stres ilişkisinde kontrol odağının rolü. [Framing Focus of Control & Workaholism Positively with Reference to Perceived Stress]. *Organizasyon ve Yönetim Bilimleri Dergisi*, *2*(1), 47-55.

Bichi, A.A., Embong, R, Mamat, M & Maiwada, D.A. (2015). Comparison of classical test theory and item response theory: A Review of empirical studies. *Australian Journal of Basic and Applied Sciences*, 9 (7), 549-556.

Bichi, A.A. & Talib, R. (2018). Item response theory: An introduction to latent trait models to test and item development. *International Journal of Evaluation and Research in Education* (IJERE), *7*(2), 142-151.

Bilgen, Ö.B., & Doğan, N., (2017). Çok kategorili parametrik ve parametrik olmayan madde tepki kuramı modellerinin karşılaştırılması [Comparison of polytomous parametric and nonparametric item response theory models]. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(3), 354-372.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: *Issues and insights. Multivariate Behavioral Research*, *36*, 523-562.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. USA: Rinehart and Winston Inc.

Demars, C. (2010). *Item response theory. Understanding statistics measurement*. (Turkish translation editor: H. Kellecioğlu). Oxford University Press

Doğan, N. & Tezbaşaran, A. (2003). Klasik test kuramı ve örtük özellikler kuramının örneklemler bağlamında karşılaştırılması [Comparison of classical test theory and latent traits theory by Samples]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 25(25)*, 58-67.

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

Erci, B. (2006) Reliability and validity of the Turkish version of perceived stress scale. *Atatürk Üniv. Hemşirelik Yüksekokulu Dergisi,* 9 (1), 52-67.

Erkuş, A., Sünbül, Ö., Sünbül, S., Yormaz, S., & Aşiret, S. (2017) *Psikolojide ve Ölçme Ve Ölçek Geliştirme-II* [Testing and Scale Development in Psychology]. Ankara: Pegem Akademi

Eskin, M., Harlak, H., Demirkıran, F., & Dereboy, Ç. (2013). Algılanan stres ölçeğinin Türkçe'ye uyarlanması: Güvenirlik ve geçerlik Analizi [The adaptation of the perceived stress scale into Turkish: A reliability and validity analysis]. *New/Yeni Journal*, *51* (3), 132-140

Fan, X. T. (1998). Item response theory and classical test theory: an empirical comparison of their item / person statistics. *Educational and Psychological Measurement, 58(3), 357-381.*

Gelbal, S., (1994). p madde güçlük indeksi ile Rasch modelinin b parametresi ve bunlara dayalı yetenek ölçüleri üzerine bir karşılaştırma [p parameter of the Rasch model with the item difficulty index and A comparison of measures based on ability]. *Hacettepe University Journal of Education Faculty*, *10*, 85-94.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston: Academic Puslishers Group

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory.* Sage Publications, London.

Hambleton, R.K., & Jones, R.W.(1993) Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12* (3), 3847.

Kan, A. (2006). Klasik test teorisine ve örtük özellikler teorisine göre kestirilen madde parametrelerinin karşılaştırılması üzerine ampirik bir çalışma. [An empirical study on the comparison of predicted item parameters with respect to classical and item response test theories]. *Mersin Üniversitesi Eğitim Bilimleri Dergisi, 2* (2), 227-235.

Karaca, A., Yıldırım, N., Ankaralı, H., Çıkgöz, F., & Akkuş, D. (2015). Hemşirelik öğrencileri için algılanan stres, biyo-psiko-sosyal cevap ve stresle başetme davranışları ölçeklerinin Türkçe'ye uyarlanması [Turkish adaptation of perceived stress scale, Bio-psycho-social response, and coping behaviours of stress scales for nursing students]. *Psikiyatri Hemşireliği Dergisi*, *6* (1), 15-25.

Köksal, G., ve Kabasakal, Z. (2012) Zihinsel engelli çocukları olan ebeveynlerin yaşamlarında algıladıkları stresi yordayan faktörlerin incelenmesi. [The examination of predicting factors of perceived stress of parents with mental retarded children]. *Buca Eğitim Fakültesi Dergisi*, *32*, 71-91.

Köse, A. (2015). Aşamalı tepki modeli ve klasik test kuramı altında elde edilen test ve madde parametrelerinin karşılaştırılması. [Comparison of test and item parameters under graded response model (IRT) and classical test theory]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 15*(2), 184 - 197.

LaHuis, D. M., & Copeland, D. A. (2009). Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organizational Research Methods, 12*, 396-319.

Lautenschlager, G. J., Meade, A. W., & Kim, S. H. (2006). Cautions regarding sample characteristics when using the graded response model. Paper presented at the 21" Annual Conference of the *Society for Industrial and Organizational Psychology*, Dallas, Texas

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

Lord, F. M. (1980). *Aplications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Nartgün, Z. (2002). *Aynı tutumu ölçmeye yönelik Likert tipi ölçek ile metrik ölçeğin madde ve ölçek özelliklerinin klasik test kuramı ve örtük özellikler kuramına göre incelenmesi.* yayımlanmamış doktora tezi [The investigation of item and scale properties of Likert type scale and metric scale measuring the same attitude according to classisical test theory and item response theory], Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Otrar, M., Ekşi, H., Dilmaç, B. & Şirin, A. (2002). Türkiye'de öğrenim gören Türk ve akraba topluluk öğrencilerinin stres kaynakları, başa çıkma tarzları ile ruh sağlığı arasındaki ilişki üzerine bir araştırma. [The sources of stress, coping, and psychological Well-Being among Turkic and relative societies' students in Turkey]. *Kuram ve Uygulamada Eğitim Bilimleri*, *2* (2), 473-506

Reckase, M.D. (2007). *Multidimensional item response theory*. C.R. Rao, S. Sinharay, (Ed.) *Handbook of Statistics, Vol, 26: Psychometrics* (pp. 607-642) Amsterdam: Elsevier

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

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance*, *14*, 187-207.

Şanlı, Ö. (2017). Öğretmenlerin algılanan stres düzeylerinin çeşitli değişkenler açısından incelenmesi [Analysing the Perceived Stress Level of Teachers with Regards to Some Variables]. *Electronic Journal of Social Sciences*. *16*(61), 85-96.

Savcı, M., & Aysan, F. (2014). Üniversite öğrencilerinde algılanan stres düzeyi ile stresle ile başa çıkma stratejileri arasındaki ilişki [The Relationship Between the Perceived Stress Level and the Stress Coping Strategies in University Students]. *Uluslararası Türk Eğitim Bilimleri Dergisi*. 3, 44-56.

Sijstma, K. & Junker, B.W. (2006). Item response theory: Past performance, present developments and future expectatitons. *Behaviormetrika, 33*(1), 75-102.

Sijtsma, K., & Meijer, R.R. (2007). *Nonparametric item response theory and special topicd.* In C.R. Rao and S. Sinharary (Eds.) *Handbook of Statistics,* Vol, 26: Psychometrics (pp. 719-746) Amsterdam: Elsevier

Stage, C. (1998a). A Comparison between item analysis based on item response theory and classical test theory. A study of the SweSAT Subtest WORD. Report, Sweden Umea

University, Department of Educational Measurement. [Online]: Retrieved on 04-December-2007, at URL: http://www.umu.se/edmeas/publikationer/pdf/enr2998sec.pdf

Sünbül, Ö. & Erkuş, A. (2013) Madde parametrelerinin değişmezliğinin çeşitli boyutluluk özelliği gösteren yapılarda madde tepki kuramına göre incelenmesi. [Examining item parameter invariance for several dimensionality types by using unidimensional item response theory]. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 9*(2), 378-398.

Tatar, A, Saltukoğlu, G. & Özmen, E. (2018) Madde yanıt kuramıyla öz bildirim türü stres ölçeği geliştirme çalışması – I: Madde seçimi, faktör yapısının oluşturulması ve psikometrik özelliklerinin incelenmesi. [Development of a self report stress scale using item response theory-I: Item selection, formation of factor structure and examination of its psychometric properties]. *Arch Neuropsychiatry, 55,* 161−170. https://doi.org/10.5152/npa.2017.18065

Turgut, M.F., (1992). *Eğitimde Ölçme ve Değerlendirme Teknikleri*. [Assessment and Evaluation in Education]. Ankara: Saydam Matbaacılık.

Turgut, M.F., & Baykul, Y. (1992). *Ölçekleme Teknikleri*. [Scaling Techniques]. ÖSYM Yayınları. Yayın No 1. Ankara

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement*, *8*, 125-145.

APPENDIX

Percieved Stress Scale Items

Items	1-Never 2-Rarely 3-Sometimes 4-Often 5-Always					
1 (2)	I feel like stress is a part of my life	①	②	③	④	⑤
2 (3)	I often feel unnecessarily over-stressed	①	②	③	④	⑤
3 (4)	I usually feel that I am an angry person	①	②	③	④	⑤
4 (5)	I usually feel very nervous because of the things I want to do but can't.	①	②	③	④	⑤
5 (6)	I feel like I'm too hasty on many things.	①	②	③	④	⑤
6 (7)	I feel that when I feel distressed, I'm not successful at comforting myself.	①	②	③	④	⑤
7 (8)	I generally feel mentally tired/exhausted.	①	②	③	④	⑤
8 (10)	I generally feel sad.	①	②	③	④	⑤
9 (11)	The feeling of not being able to control the disorder in my life makes me angry.	①	②	③	④	⑤
10 (12)	The thought that I can't control my anger sometimes, scares me.	①	②	③	④	⑤
11(13)	The feeling that I won't be able to overcome the problems that I'm facing bothers me	①	②	③	④	⑤
12 (28)	I'm very worried about the extreme responsibilities I've been given.	①	②	③	④	⑤
13 (29)	Sometimes I think that the works I'm going to take on are excessive.	①	②	③	④	⑤
14 (32)	The feeling that others' expectations of me are too extreme, bothers me.	①	②	③	④	⑤
15 (35)	The possibility of making mistakes in extreme decisions that I will make in life makes me avoid making decisions.	①	②	③	④	⑤
16 (45)	I always feel mentally tired/exhausted	①	②	③	④	⑤

Numbers in parentheses indicate the number of questions on the draft scale