ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2016; 7(1);133-144

The Development of IRT Based Attitude Scale towards Educational Measurement Course*

Ölçme ve Değerlendirme Dersine Yönelik Tutum Ölçeğinin Madde Tepki Kuramına Dayalı Olarak Geliştirilmesi*

Nükhet DEMİRTAŞLI ** Seher YALÇIN *** Cansu AYAN ****

Abstract
In this study, the Scale of Attitude towards Educational Measurement and Evaluation (SAEM) developed by Demirtaşlı (2002) is reconstructed based on polytomous Item Response Theory (IRT) models and its psychometric features are identified. In this context, the best polythomous IRT model was investigated which is fitted SAEM data. IRT models gives invariant person and item parameters, when data-model fit. A version of SAEM has 41 Likert type items with four points was administered to 519 teacher candidates attending teacher education programs at several universities in Turkey. The data were analyzed according to polythomous IRT models: Samejima's graded response model (S-GRM), the partial credit model (PCM) and a nominal response model (NRM). The results of the analysis showed that a new version of SAEM, which is based on S-GRM, consists of 33 items, has lower chi-square value than the other models and the classic internal reliability was found to be 0.97. The findings of the study indicate that the validity and reliability features of the scale are fairly good.

Key Words: Attitude toward educational measurement and evaluation, polytomous item response model, attitude scale.

Öz
Bu araştırmada, ölçme ve değerlendirme dersine yönelik tutumu ölçmek üzere geliştirilen Likert tipi, Ölçme ve Değerlendirme Dersine Yönelik Tutum (ÖDET) ölçeğinin (Demirtaşlı, 2002) madde tepki kuramına dayalı (MTK) çok kategorili puanlanan modeller çerçevesinde yeniden ölçeklenerek psikometrik özelliklerinin karşılaştırılması amaçlanmıştır. MTK'ya dayalı modeller verilerle uyum gösterdiğinde, değişmez birey ve madde parametreleri kestirilebilir amaca uygun ölçek geliştirmede daha güvenilir ve geçerli sonuçlara ulaşılabilir. Ölçeğin ilk versiyonu, dörtlü Likert tipi, dereceli toplamlı tepki vermeye uygun 41 maddeden oluşmaktadır. Ölçeğin bu formu, Türkiye'deki farklı illerden üç devlet üniversitesinin öğretmen yetiştiren fakülte ve programlarına devam eden 519 üniversite öğrencisine uygulanmıştır. Maddeler çok kategorili puanlanan maddeler için geliştirilen MTK modellerinden Samejima (S) kademeli tepki modeli (S-Graded Response model), kısmi puan modeli (partial credit model) ve sınıflandırmalı tepki modeline (nominal response model) göre ölçeklenmiştir. Ki-kare veri-model uyum testi sonucunda, S-kademeli tepki modeline göre ölçeklenen ölçeğin 33 maddelik yeni versiyonunun veriyle daha uyumlu olduğu görülmüştür. İç tutarlık anlamındaki klasik güvenilirlik katsayısının da 0.97 olduğu bulunmuştur. S-kademeli tepki modeline göre psikometrik özellikleri MTK'ya göre kestirilen ÖDET'in geçerli ve güvenilir bir ölçme aracı olduğu görülmüştür.

Anahtar Kelimeler: Ölçme ve değerlendirmeye yönelik tutum, çok kategorili MTK, tutum ölçeği.

* Paper presented at The 4th Congress on Measurement and Evaluation in Education and Psychology on June 9th-13th 2014, Ankara

** Prof. Dr., Ankara University, Faculty of Educational Sciences, Ankara-Turkey, e-mail: rnukhet@yahoo.com
*** Res. Asst., PhD., Ankara University, Faculty of Educational Sciences, Ankara-Turkey, e-mail: yalcins@ankara.edu.tr
**** Res. Asst., Ankara University, Faculty of Educational Sciences, Ankara-Turkey, e-mail: cnsayan@gmail.com

DOI: 10.21031/epod.43804
Geliş Tarihi: 17.11.2015
Kabul Tarihi.17.05.2016

INTRODUCTION

Educational measurement and evaluation is a compulsory course in undergraduate and teacher certification programs. In spite of, teachers spend as much as a third of their professional time in assessment related activities and many of these activities require skills in testing and measurement (Wise, Lukin and Roos, 1991), some pre-service and in-service teachers have concerns and negative attitude about succeeding in these math-based subjects (Brady & Bowd, 2005; Gresham, 2010; Jaggernouth, 2010; Kottke, 2000). As an affective trait, attitude is a tendency to respond in direction of approaching or avoiding to an object, person, institution, or event (Ajzen, 2005). This tendency can have an indirect positive or negative impact on learning behavior (Perkins, Adams, Pollock, Finkelstein & Wieman, 2005; Reed, Drijvers & Kirschner, 2010; Shih & Gamon, 2001). Several studies have investigated the attitudes of pre-service and in-service teachers towards the measurement and evaluation course and their self-efficacy in this course (Aktaş & Alıcı, 2012; Kılınç, 2011; Kilmen & Demirtaşlı, 2009; Ozan & Köse, 2013; Özbaşı & Demirtaşlı, 2013; Ulutaş, 2003). Recognizing the attitudes of pre- and in-service teachers towards the measurement and evaluation course can be used to create a more positive learning environment in education and training programs. Searching and analyzing the negative attitudes of student teachers in relation to a course using a valid and reliable attitude scale helps to identify the pedagogic action to be taken to change teacher candidates's attitudes from negative to positive. This situation contributes to the establish a positive learning climate.

In education and psychology, test construction is based on primarily two test theories; the classical test theory (CTT) and the item response theory (IRT). The theoretical foundations of IRT dates back to the 1950s however, since IRT-based estimations require complex mathematical and statistical processes, the remarkable progress in this area was observed after the 1980s with the significant innovations in computer and software technology. When studied on a dataset that meets its basic assumptions, IRT can overcome the limitations of CTT and provides several advantages for the scaling process. In scale-development studies based on IRT, when the basic assumptions of IRT are fulfilled and the data fit the model, invariant person and item parameters can be estimated (De Ayala, 2009; Hambleton et al., 1991). Therefore, IRT based tests are not necessarily to establish conventional test norms for items measure in the same way at subsamples from the same population (Embretson & Reise, 2000, p. 25; Hambleton et al., 1999). An IRT-based scale can be used as a valid and reliable instrument to estimate the traits of subsamples. With this advantage, IRT can also be used to solve other measurement problems such as those related to the test equating, computer based adaptive testing, detecting of biased items.

In this context, the purpose of IRT based SAEM is to benefit from IRT's advantages such invariant item and theta parameters when model-data fit. By means of invariance, no further norm studies in interpretation of SAEM scores, comparison of groups. Besides, since IRT models give individual error estimations in item and person level, IRT based SAEM will be able to measure attitudes towards educational measurement course more reliably. In addition to this advantage, it can be detect possible item bias for several participants' background variables like type of under graduate program (Social sciences, Science), level of attitudes towards numerical content courses. Finally, when SAEM developed based on IRT, paralell forms of SAEM can be construct more easily and reliably.

Purpose

In this study, IRT was used to reconstruct a Likert-type CTT-based scale (SAEM) developed by Demirtaşlı (2002) to measure the attitude towards educational measurement and evaluation. In this context, the best polythomous IRT model that fits attitude data was investigated. For this purpose, the psychometric characteristics of the SAEM were tested under Samejima's Graded Response Model (GRM), Partial Credit Model (PCM) and Nominal Response Model (NRM) (Embretson & Reise, 2000)

METHOD

Research Model

This study is a descriptive research since the aim was to identify the psychometric features of the SAEM based on IRT (Glass & Hopkins, 1984; Kaptan, 1995).

Study Group

This scale was administered to 519 pre-service teachers enrolled in teacher college in three public universities in three different provinces of Turkey. All the participants had already taken the measurement and evaluation course in teacher college programs. Of the participants, 67% were female and 29% were male. Participation in the study was voluntary.

Data Collection

The scale reconstructed in this study was developed to measure the attitudes towards the measurement and evaluation course, which is compulsory in teacher education and teacher certification programs. This scale is a four-point graded Likert scale consisting of 41 items, and was found to measure valid and reliable with three factors. The results of Cronbach's alfa correlation coefficients showed that the reliability for SAEM's each factor were ranged from .82 and .92 (Demirtaşlı, 2002). The following four categories was used to respond to all items in the scale; 1=strongly disagree, 2=disagree, 3=agree and 4=strongly agree. The items were scored reverse that are express negative attitudes: 4 for "strongly disagree" and 1 for "strongly agree". The minimum and maximum scores of the scale are 41 and 164, respectively. A higher score means that the participant has a more positive attitude towards the measurement and evaluation course, and a lower score indicates a negative attitude.

Data Analysis

Data analysis was performed in three stages. First, the participants' responses to the items were scored. Then, the data were analyzed in terms of basic IRT assumptions namely unidimensionality and local independence. When the data fit the IRT-based models, invariant person and item parameters can be estimated (Embretson & Reise, 2000; Hambleton et al., 1991). This feature of IRT helps to construct tests for the expected features, and also, equate test forms and develop computerized adaptive testing.

The scale dimensionality was detected by a Principal Component Analysis (PCA). The data were analyzed by the SPSS 15.0. The statistical convenience of the items to the PCA was determined using their Kaiser-Mayer-Olkin (KMO) value and the results obtained from Bartlett's test. The KMO value was found to be 0.97, and according to the result of Bartlett's test, the chi-square statistic was significant ($\chi^2(820) = 13163.31$; $p<0.05$). These findings indicate that the items of the scale fit the PCA. In the first analysis, 41 items were loaded under five components. In initial analysis, the five-component structure was observed that accounted for 60% of the total variance and ten items had loadings more than one factor (factor loading > 0.40). The scree plot (Figure 1) of the data shows a rapid decrease in the eigenvalue from the first to the second factor. Based on this result, it was concluded that SAEM had a dominant one factor.

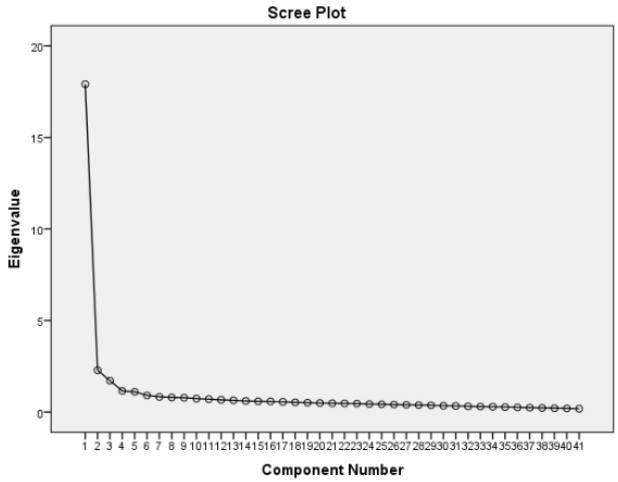

Figure 1. Scree Plot of the SAEM Factor Structure

After that, factor structure of the scale was re-analyzed by restricting it to a single factor with varimax rotation. The results of the PCA restricted to a single factor showed that 41 items explained 44% of the total variance and factor loadings were varied from 0.35 to 0.77. Based on these results, it can be concluded that SAEM has a dominantly unidimensional structure and thus met the unidimensionality assumption of IRT. Another assumption of IRT is local independence, which means that at a given trait level, a test taker's response to an item is independent from the other items. In other words, a response to any of the items in the scale (e.g. endorsing or rejecting a certain attitude) is not dependent on the response to another item. This is observed when the unidimensionality assumption is met. In a test identified as unidimensional, the covariance between the items is zero for subjects at the same latent trait. This indicates that once the unidimensional assumption is met, the local independence assumption is also met (Hambleton & Swaminathan, 1985). As a result, the 41-item scale used in the current study was considered to have fulfilled the assumption of local independence.

In the second phase of the data analysis, items were detected in terms of bias. The probable source of bias for this data is gender. In testing procedure, the individual differences resulted from the measured trait, rather than the gender of the participants with the same latent trait. To this end, the items in the scale were analyzed to determine whether they displayed differential item functioning (DIF) in terms of gender. To detect the DIF of polytomous items, the Polytomous Simultaneous Item Bias Test (PSIBTEST) and IRT Likelihood Ratio Test (IRT-LRT) were used. In the PSIBTEST method, DIF is determined through a regression-based correction that can determine Type I error (Clauser & Mazor, 1998).

IRT-LRT is based on a comparison of observed and theoretical models (Thissen, Steinberg & Wainer, 1993) which requires restricted and extended models. In the restricted model, which assumes that none of the items has DIF, the probability of the parameters of all items being equal is calculated. In the extended model, the likelihood of item parameters, for which DIF is detected, being different in the focal and reference groups with other parameters being equal is found. The G^2 value is calculated by subtracting the two −2log likelihood values obtained from the likelihood ratio of the restricted and extended models (Thissen, 2001). The calculated G^2 value is then compared to the chi-square value with the degrees of freedom. The degrees of freedom is the number of parameters in the model, and thus in the current study, it was four (df= three threshold parameters and one discrimination parameter). If the G^2 value is less than 9.49, it is interpreted that a negligible DIF level is present; if it is higher, then there is a medium or high level of DIF against the focal

group of the relevant item (Greer, 2004). The IRT-LRT analysis method uses anchor items to equate the groups. For the selection of anchor items, the following criteria are used; having a high level of discrimination, having a high range of difficulty level, displaying no DIF according to other DIF detection methods, producing a small error variance in the PCA and having high factor loadings (Yıldırım, 2006). In this study, the criteria for the selection of anchor items were that they represented both way of the attitude, display no DIF according to the result of PSIBTEST and have high factor loadings in PCA. As a result, items 19, 27, 29 and 34 were selected as anchor. DIF analyses were performed DIFPACK 1.7 and IRTLRDIF 2.0b packages. Table 1 presents the results of the analyses performed using two DIF detection methods.

Table 1. Results of the DIF Analyses Under Two Different DIF Methods

	PSIBTEST	**IRT-LRT**
	B or C Level DIF	**B or C Level DIF**
Items	3, 8, 10, **13, 17,** 18, 19, 25, 26, 31, 37	**13, 17,** 41

As shown in Table 1, two items were found to display DIF in both methods (items 13 and 17). Item 13 was, "I would like to take other measurement courses" and item 17 was, "I wish I could take more measurement and evaluation courses". Both items were in favor of the male participants. In other words, when male and female students with the same level of attitude were compared, the probability of male students moving from "agree" to "strongly agree" was found to be significantly higher. Following the analysis performed, these items were excluded from the scale.

In the third stage of data analysis, the remaining 39 items were analyzed according to Samejima's GRM, PCM and NRM using the MULTILOG 7.03 package. Samejima's GRM is used to measure items that have ordered categorical responses such as Likert type scale items, and it is an extension of the two-parameter logistic (2PL) model. In GRM, the threshold values of response categories should be ordered, which is not required by PCM or generalized PCM (Embretson & Reise, 2000). PCM was originally developed for items that require responses in multiple steps. It is also used for the analysis of responses to items in scales that measure traits, in which two or more categorical responses are possible (such as attitude and personality traits). NRM is used to measure responses of similar format items but it does not require item choices to be ordered or identified numerically. The purpose of this model is to plot options characteristic curves based on the frequency of the selected choices in multiple-choice items. This model can also be applied to attitude and personality scales. All three models are used in items that are scored using grades and they have different advantages and disadvantages. For example, Samejima's GRM does not require the items to have the same number of categories. Therefore, it is appropriate for scales consisting of items with different response formats. Furthermore, this model is an extension of the 2PL model and allows the discrimination index to be different among items. PCM, on the other hand, is an extension of the Rasch Model, and as a result, raw score is sufficient statistics to estimate the ability level of an individual. However, in the PCM model, the slopes of all items in this model are considered to be equal. In other words, the model assumes that the discrimination index among items is equal, which is not that easy to realize in practice (Baker, Rounds & Zevon, 2000; Embretson & Reise, 2000).

RESULTS

Thirty-nine items of the scale were scaled using the three models, and Table 2 presents the maximum item information obtained from each model.

Table 2. Amount of Information Obtained From Items Using Different Polythomous IRT Models

Items	GRM	PCM	NRM	Items	GRM	PCM	NRM
M1	0.982	0.876	0.963	**M21**	1.441	1.992	1.557
M2	1.052	1.135	1.357	**M22**	0.982	1.658	1.028
M3	0.577	0.774	0.596	**M23**	1.254	1.731	1.171
M4	0.825	1.109	0.927	**M24**	1.250	1.814	1.246
M5	1.561	2.075	1.540	**M25**	2.135	3.568	2.386
M6	0.388	0.394	0.684	**M26**	1.470	1.942	1.306
M7	0.351	0.469	0.300	**M27**	2.749	4.418	3.161
M8	0.491	0.688	0.477	**M28**	0.880	1.404	1.112
M9	0.784	0.849	0.836	**M29**	0.138	0.147	0.134
M10	0.231	0.246	0.222	**M30**	1.424	1.686	1.578
M11	0.533	0.630	0.551	**M31**	1.684	3.321	2.201
M12	1.886	3.222	2.022	**M32**	2.891	7.068	3.011
M13	0.792	0.910	0.939	**M33**	0.825	0.923	0.961
M14	1.451	1.916	1.373	**M34**	1.996	4.063	2.439
M15	1.754	2.779	1.825	**M35**	1.669	2.497	1.921
M16	0.445	0.483	0.631	**M36**	1.822	2.976	2.076
M17	1.773	2.198	1.727	**M37**	1.760	2.469	1.979
M18	1.868	3.008	2.115	**M38**	1.250	2.104	1.196
M19	1.804	3.617	1.739	**M39**	0.441	0.487	0.520
M20	1.931	3.075	3.104				
Total Information	**49.84 (-1.40)***	**67.56 (-1.20)**	**45.97 (0.60)**	**Marginal Reliability**	**0.973**	**0.974**	**0.970**

*The values in parentheses indicate the level of trait (attitude) with the highest amount of information.

As shown in Table 2, according to GRM, PCM and NRM, item information ranges from 0.14 to 2.89, from 0.15 to 7.07 and from 0.13 to 3.10, respectively. The total test information values obtained from the three models are presented in Table 2. The highest test information (67.56) provided from PCM at -1.20 theta (attitude) level. The highest test informations obtained from GRM and NRM respectively. Although the reliability coefficient of all three models was close to each other, the highest reliability coefficient value, 97.4, was obtained from PCM.

The model-data fit level was determined by comparing -2 log likelihood values from polythomous model pairs. First, PCM and GRM were compared in terms of differences in -2 log χ^2 values, chi-square statistics and degrees of freedom (Df). Df is computed by multiplying the number of items with the number of parameters calculated for the estimation model. The number of parameters varies according to the model used for estimation; however, the number of "step difficulty/threshold/intercept" parameters substituting item difficulty equals the "number of categories-1" (Embretson & Reise, 2000). In PCM, for each item with four categories, three step difficulty parameters and two item slope parameters were calculated, and thus the degrees of freedom is 195 (39*5). In GRM, for each item with four categories, three threshold parameters and one item slope parameter were estimated, resulting in a degrees of freedom of 156 (39*4). According to this, χ^2 (195, 156)= 26115.8- 25886.5= 229.3 and the approximate table value is χ^2 (39; 0.05) = 55.75. Since the calculated value is higher than the table value, the difference between the -2 log χ^2 values is significant. Therefore, it can be concluded that GRM is more appropriate for this type of data. In the second stage, the difference in -2 log χ^2 values obtained from GRM and NRM was determined and compared with the Chi-Square statistic using the 0.05 significance level and related degrees of freedom. In NRM, for each item with four categories, three intercept and three item slope parameters are computed, which results in degrees of freedom being 234 (39*6). χ^2 (156, 234)= 25886.5 - 25832.4= 54.1 and the approximate table value is χ^2 (78; 0.05) = 101.88. Since the calculated value is lower than the table value, the difference between the -2 log χ^2 values is not significant. This indicates that there is no difference between the GRM and NRM models. Furthermore, in GRM, the reliability and maximum information values were found to be 0.973 and 49.84, respectively while in NRM, these were 0.970 and 45.97, respectively. Although no significant difference was observed between GRM and NRM in terms of data fit, estimations were performed using GRM since the highest maximum information and reliability was achieved with this model. Through the estimations on the Multilog program, the slope, threshold parameters and threshold

information functions of all the items were obtained and analyzed. According to GRM, six of the 39 items (6, 7, 10, 18, 31 and 41) had a low level of item information (under .45), and therefore, they were excluded from the scale. Furthermore, differences in the observed and expected values were analyzed for each category and the items were found to be fit to the data. Table 3 presents the slope and threshold parameters of items estimated by the GRM model (33 items).

Table 3. Estimated Item Parameters According to GRM

Items	a	b1	b2	b3	Items	a	b1	b2	b3
M1	1.84	-2.03	-1.45	0.18	**M18**	1.89	-2.23	-0.94	1.23
M2	1.91	-1.91	-1.39	0.44	**M19**	2.10	-1.67	-0.59	1.22
M3	1.37	-1.98	-0.78	1.48	**M20**	2.06	-1.73	-0.82	1.08
M4	1.74	-1.89	-0.14	1.95	**M21**	2.72	-1.69	-0.93	0.84
M5	2.36	-1.54	-0.50	1.26	**M22**	2.32	-1.57	-0.39	1.40
M6	1.28	-2.83	-1.17	1.10	**M23**	3.11	-1.47	-0.67	0.99
M7	1.65	-2.36	-1.29	0.84	**M24**	1.83	-1.81	-0.33	1.67
M8	1.38	-1.62	0.04	2.62	**M25**	2.31	-1.42	-0.01	1.67
M9	2.52	-1.92	-1.12	0.83	**M26**	2.55	-1.70	-0.47	1.35
M10	1.72	-1.97	-0.30	1.75	**M27**	3.27	-1.66	-1.13	0.83
M11	2.22	-1.60	-0.78	0.87	**M28**	1.77	-1.71	0.17	1.84
M12	2.61	-2.20	-1.15	0.73	**M29**	2.76	-1.98	-1.12	0.81
M13	2.59	-1.30	-0.07	1.67	**M30**	2.48	-1.63	-0.66	0.97
M14	2.59	-1.73	-1.08	0.67	**M31**	2.52	-1.82	-1.06	0.81
M15	2.53	-2.17	-1.37	0.62	**M32**	2.66	-2.24	-0.98	0.72
M16	2.72	-1.77	-0.78	0.96	**M33**	2.20	-2.45	-1.22	0.92
M17	2.34	-1.62	-0.29	1.57					

As shown in Table 3, parameter 'a' is between 1.28 and 3.27. DeMars (2010) stated that the discrimination level of polytomous items is interpreted in the same way as dichotomous items. The discrimination level of items is classified as; very low (0.01-0.34), low (0.35-0.64), medium (0.65-1.34), high (1.35-1.69) and very high (>1.70) (Baker, 2001). On this basis, 33 items in the current scale had a high or very high level of discrimination, with item 27 having the highest and item 6 having the lowest level.

The threshold values of categories vary from -2.83 to 2.62. Most of the first two threshold parameter values were found to be negative, which indicates that the responses to the first three categories had been endorsed by participants with much lower attitude levels ($\theta<0$). The category threshold values in Table 3 show that the first threshold parameter is around "-2", the second was around "-1" and the third was around "0". This indicates that the scale better differentiates people with a lower attitude. Furthermore, the threshold parameter increasing in parallel to the attitude level suggests that the categories performes hierarchically as expected. Considering all these results, it is concluded that the discrimination characteristic and threshold values of the items in the scale are sufficiently high. Figure 2 presents the standard errors related to total information and measurement error obtained from the 33-item GRM-based SAEM.

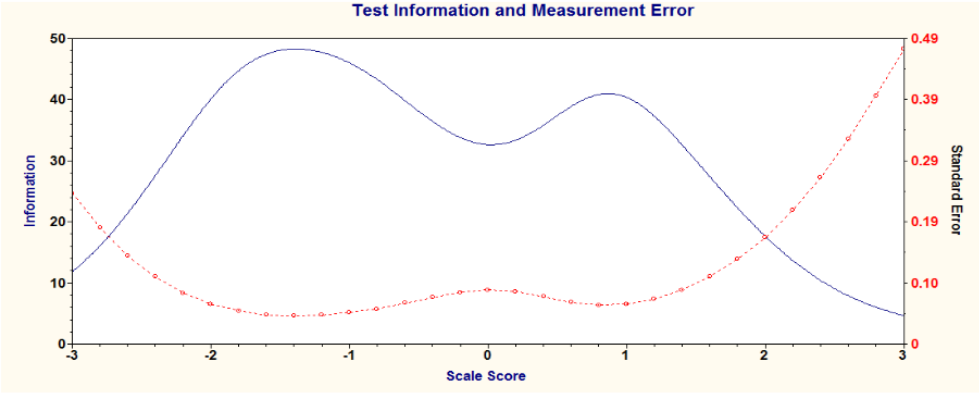

Figure 2. Standard Errors and the Total Test İnformation Obtained from the GRM-based SAEM

Figure 2 shows that in the last version of the 33-item GRM-based SAEM, the information obtained from the scale is considerably high and falls within a wide range of attitude levels ($-2 \leq \Theta \leq +2$). The scale only provided less information for individuals with an attitude level at the lowest or highest end. The maximum information obtained from the GRM-based scale was found to be 48.23, which was achieved at the attitude level of -1.40.

CONCLUSION and DISCUSSION

In related literature, there are many examples of cognitive test construction studies under IRT models. On the other hand, scales on affective traits have relatively limited that developed by IRT based models. In this study, As an attitude scale SAEM was rescaled based on polyhomous IRT models. Attitude is remarkable affective trait which is effective on our behaviours (Ajzen, 2005). Searching and displaying the negative attitudes of pre-service teachers in relation to a course using a valid and reliable attitude scale helps to identify the pedagogic efforts and activities to be taken to change negative attitudes in direction of positive attitudes. This contributes to the establishment of a positive learning environment. Differently from the CTT based scales, tests and scales based on IRT models separately estimates the probability of individuals at different trait (attitude) levels endorsing each category in each item, which provides more valid and reliable results in terms of the measures of individual differences. Thus, the functionality of both items and response categories can be estimated independently from the other items in the scale and the participant samples. In other words, from the total information and item information values obtained with this approach, it is possible to identify items and response categories that reveal the individual differences at different attitude levels (Le, 2013; Matteucci & Stracqualursi, 2006). In the current study, the model-data fit was higher in GRM than PCM. This may have resulted from the GRM criterion that the threshold values of response categories should be ordered. Thus, the statements "strongly disagree", "disagree", "agree" and "strongly agree" can be considered to indicate ordered responses. This study shows that the IRT based SAEM with 33-item is a valid and reliable instrument to determine the attitudes of pre-service teachers towards the measurement and evaluation course.

Future studies will be able to carry out with this instrument. This IRT based version of the scale is a valid and reliable scale that can be used in studies that compare the attitudes of teacher candidates from teachers and non-teachers college programs. And also, this scale can be used to investigate the effectiveness of a variety of actions undertaken during the teaching of the course to change any negative attitudes held by the pre-service teachers. Furthermore, this scale can be used to investigate the relationship between pre-service teachers' achievement in the measurement and evaluation course and their attitude towards it.

REFERENCES

Aktaş, M., & Alıcı, D. (2012). Development of likert type attitude scale towards measurement and evaluation in education course. *Journal of Qafqaz University*, *33,* 66-73.

Ajzen, I. (2005). *Attitudes, personality and behavior*. Milton-Keynes, England: Open University Press/ McGraw- Hill.

Baker, F. B. (2001). *The basis of item response theory*. USA: ERIC Clearing house on Assessment and Evaluation.

Baker, J. G., Rounds, J. B. & Zevon, M. A. (2000). A comparison of graded response and rasch partial credit models with subjective well-being. *Journal of Educational and Behavioral Statistics, 25,* 253–70.

Brady, P. & Bowd, A. (2005). Mathematics anxiety, prior experience and confidence to teach mathematics among pre-service education students. *Teachers and Teaching: Theory and Practice, 11*(1), 37–46.

Clauser, B. E.,& Mazor, K. M. (1998). Using statistical procedure to identify differential item functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31-44.

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

DeMars, C. (2010). *Item response theory*. Oxford: Oxford University Press.

Demirtaşlı, Ç. N. (2002). Developing a scale for attitudes toward the measurement and evaluation course. *Education and Science, 27*, 125, 44-48.

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

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice Hall.

Greer T. G. (2004). *Detection of differential item functioning (DIF) on the SATV: A comparison of four methods: Mantel-Haenszel, logistic regression, simultaneous item bias and likelihood ratio test* (Doctoral dissertation, University of Houston, Texas). Retrieved from http://search.proquest.com/pqdtglobal/docview/305196900/EE5E5B40D7A34E5DPQ/1?accountid=8319

Gresham, G. (2010). A study exploring exceptional education pre-service teachers' mathematics anxiety. *IUMPST: The Journal*, *4*, 1-14.

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

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

Jaggernauth, J. S. (2010). *Mathematics anxiety and the primary school teacher: An exploratory study of the relationship between mathematics anxiety, mathematics teacher efficacy, and mathematics avoidance.* EDRS6900: Project Report. The University of the West Indies.

Kaptan, S. (1995). *Bilimsel araştırma ve istatistik teknikleri* (10. Baskı). Ankara: Rehber Yayınevi.

Kılınç, M. (2011). A perceptual scale for measurement and evaluation of prospective teachers self-efficacy in education. *Journal of Kırşehir Education Faculty, 12*, 4, 81-93.

Kilmen, S., & Çıkrıkçı-Demirtaşlı, N. (2009). The Perceptions of primary school teachers about their application levels of measurement and evaluation principles. *Ankara University Journal of Faculty of Educational Sciences*, *42*(2), 027-054.

Kottke, J. L. (2000). Mathematical proficiency, statistics knowledge, attitudes toward statistics, and measurement course performance. *The College Student Journal*, *34*, 334-347.

Le, D. T. (2013). *Applying item response theory modeling in educational research* (Doctoral dissertation, Iowa State University. Retrieved from http://search.proquest.com/pqdtglobal/docview/1450045295/DD5A94235C534A49PQ/1?accountid=8319.

Matteucci, M., & Stracqualursi, L. (2006). Student assessment via graded response model. *STATISTICA*, *4*, 435–447.

Ozan, C., & Köse, E. (2013). Adaptation of attitudes toward educational measurement inventory (ATEMI) to Turkish. *E-İnternational Journal of Educational Research*, *4* (2), 29-47.

Özbaşı, D., & Demirtaşlı-Çıkrıkçı N. (2013). Primary school teachers' perceptions of their competencies regarding measurement and evaluation in terms of some variables. *Ankara University Journal of Faculty of Educational Sciences*, *46*(2), 25-46.

Perkins, K. K., Adams, W. K., Pollock, S. J., Finkelstein, N. D., & Wieman, C. E. (2005). Correlating student beliefs with student learning using the Colorado learning attitudes about science survey. *AIP Conference Proceedings*, 790(1), 61–64.

Reed, H. C., Drijvers, P., & Kirschner, P. A. (2010). Effects of attitudes and behaviours on learning mathematics with computer tools. *Computers & Education*, 55(1), 1-15.

Shih, C. C., & Gamon, J. (2001). Web-based learning: Relationships among student motivation, attitudes, learning styles, and achievement. *Journal of Agricultural Education*, 42(4), 12-20.

Thissen, D. (2001). IRTLRDIF v.2.0b: *Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning.* Chapel Hill: L. L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland and H. Wainer (Eds.). *Differantial item functioning*. Lawrence Erlbaum.

Ulutaş, S. (2003). *Investigating the competency of teachers in high schools in measurement and evaluation and level of application principles of measurement and evaluation* (Master thesis, University of Ankara). Retrieved from https://tez.yok.gov.tr/.

Wise, S.L., Lukin, L.E. & Roos, L.L. (1991). Teacher beliefs about training in testing and measurement. Journal of Teacher Education, 42(1), 37-42.

Yıldırım, H. H. (2006). *The differential item functionıng (DIF) analysis of mathematics items in the international assessment programs* (Doctoral dissertation, METU). Retrieved from https://tez.yok.gov.tr/.

GENİŞ ÖZET

Giriş

Ölçme ve değerlendirme dersi öğretmen yetiştiren lisans programlarında ve öğretmenlik sertifikası programlarında okutulan zorunlu bir derstir. Bu dersin, genel öğretmenlik yeterliklerini geliştirmek üzere verilen temel dersler içinde matematiksel ve istatistiksel işlemlerle ilgili konuları da içermesi nedeniyle "sayısal" bir ders olarak algılanması, öğretmen adaylarının bu derste başarılı olma konusunda kaygılı olmaları, bu derse ilişkin öğretmenlerin ve öğretmen adaylarının tutumlarının bilinmesinin gerekliliği göz önünde bulundurulduğunda geçerli ve güvenilir bir ölçme aracına ihtiyaç vardır. Literatürdeki ölçme araçlarının Klasik Test Kuramına (KTK) dayalı olarak geliştirilmesi, farklı gruplarda tekrar geçerlilik ve güvenirlik kanıtları toplanmasını gerektirmektedir. Madde Tepki Kuramı'nda (MTK) ise varsayımlar sağlandığında, bir ölçme aracı aynı evrenden geldiği bilinen alt örneklemlerinde de ölçme amacına hizmet etmektedir. Bu nedenle bu çalışmada, ölçme aracının yapısına uygun MTK modeli ile ölçekleme yapılmasına gereklilik görülmüştür.

Bu araştırmada, ölçme ve değerlendirme dersine yönelik tutumu (ÖDET) ölçmek üzere Demirtaşlı (2002) tarafından geliştirilen Likert tipi tutum ölçeğinin madde tepki kuramına dayalı çok kategorili puanlanan modellerden Samejima (S) Kademeli Tepki Modeli-KTM (Graded Response Model-S-GRM), Kısmi Puan Modeli-KPM (Partial Credit Model) ve Sınıflamalı Tepki Modeli'ne-STM (Nominal Response Model) (Embretson ve Reise, 2000) göre yeniden ölçeklenip psikometrik özelliklerinin karşılaştırılarak en uygun MTK modelinin saptanması ve bu modele göre ölçeğin psikometrik niteliklerinin betimlenmesi amaçlanmıştır.

Yöntem

Betimsel türdeki bu araştırmanın katılımcıları, Türkiye'nin üç ilindeki üç devlet üniversitesinin öğretmen yetiştiren fakülte ve programlarına devam eden, ölçme ve değerlendirme dersi almış, toplam 519 öğretmen adayından oluşmaktadır.

Araştırma kapsamında kullanılan ölçeğin dörtlü Likert tipi dereceli toplamlı tepki vermeye uygun 41 maddeden oluşan ilk versiyonunda (Demirtaşlı, 2002), her bir madde dört kategoriden birinde tepkide bulunmaya uygundur (1=hiç katılmıyorum, 2=katılmıyorum, 3=katılıyorum ve 4=tamamen katılıyorum).

Bu çalışmada, verilerin analizi üç aşamada gerçekleştirilmiştir. İlk aşamada veriler, MTK'nın temel varsayımları olan tek boyutluluk ve yerel bağımsızlık bakımından incelenmiştir. Ölçeğin başat bir boyutu ölçüp ölçmediği Temel Bileşenler Analizi (TBA) ile incelenmiştir. Veriler SPSS 15.0

programı ile analiz edilmiştir. Bulgular, maddelerin TBA için uygun olduğunu göstermektedir. Analiz sonucu, ÖDET'in başat bir faktörü yokladığı sonucuna ulaşılmıştır. Bu durum, tek boyutluluk varsayımının karşılanmasıyla yerel bağımsızlık varsayımının da karşılandığını göstermektedir (Hambleton ve Swaminathan, 1985).

Verilerin analizinin ikinci aşamasında, ölçekteki maddelerin kız ve erkek öğretmen adayları arasında ölçülen özellik dışında cinsiyete göre madde işlev farklılığı (MİF) (Differential Item functioning) gösterip göstermediği incelenmiştir. Bu amaçla çoklu puanlanan maddeler için MİF'i belirlemeyi sağlayan PSIBTEST (Polytomous Simultaneous Item Bias Test) ve Madde Tepki Kuramı Olabilirlik Oran Testi (MTK-OO) (Item Response Theory Likelihood Ratio Test) yöntemleri kullanılmıştır. MİF analizlerinde DIFPACK 1.7 ve IRTLRDIF 2.0 paketleri kullanılmıştır. İki yöntemle de ortak olarak erkekler lehine MİF gösteren iki madde olduğu görülmüş ve uzman görüşleri sonucu çıkarılmasının uygun olduğuna karar verilmiştir.

Verilerin analizinin üçüncü aşamasında, kalan 39 madde MULTILOG 7.03 paket programında Samejima KTM, KPM ve STM modeline göre analiz edilmiştir. Model tercihi test bilgi fonksiyonu değeri ve güvenirlik değerine göre yapılmıştır.

Sonuç ve Tartışma

Bu araştırmada öğretmen yetiştirme programında zorunlu ders kategorisinde olan ölçme ve değerlendirme dersine yönelik tutumları ölçmek üzere geliştirilmiş olan bir aracın (ÖDET) MTK'ya göre yeniden ölçeklenerek, geliştirilmesi amaçlanmıştır. Çok kategorili veriler için geliştirilen MTK modellerine göre madde ve test parametreleri kestirilen ÖDET'in en yüksek uyum gösterdiği MTK modelini saptamak üzere, modellerden kestirilen -2 loglikelihood değerleri ikili olarak karşılaştırılmıştır. İlk olarak KPM ve KTM'ye ait -2 log $\chi 2$ değerlerinin farkı, Kay-Kare istatistiği ve serbestlik derecesi değerlendirilmiş ve veriler için Kademeli Tepki Modeli-KTM'nin daha uygun olduğu görülmüştür. İkinci aşamada, KTM ve STM'ye ait kestirimlere ilişkin -2 log $\chi 2$ değerlerinin farkı alınarak ilgili serbestlik derecesi ile tabloda verilen Kay-Kare istatistiği ile karşılaştırılmıştır. Modeller arasında bir farklılık olmadığı tespit edilmiştir. Analizler sonucu, KTM ve STM'nin veriye uyumu açısından manidar farklılık olmamakla beraber KTM'nin maksimum bilgi ve güvenirliğinin daha yüksek olması sebebiyle, kestirimler bu modele göre yapılmıştır. KTM'ye göre yapılan incelemeler sonucu, 39 maddeden altısının (6, 7, 10, 18, 31, 41 nolu maddeler) madde bilgi düzeylerinin düşük olduğu (.45'in altında) görülmüş ve ölçekten çıkarılmıştır. Ayrıca her bir kategori için gözlenen ve beklenen oranlar arasındaki farklar da incelenmiş, maddelerin veriye uyumlu olduğu görülmüştür.

KTM ile elde edilen sonuçların kısmi puanlama modelinden daha fazla uyum göstermesi, Kademeli Tepki Modeli'nde tepki kategori eşik değerlerinin sıralı olması koşulunun bulunmasından kaynaklı olabilir. Bu durum "hiç katılmıyorum, katılmıyorum, katılıyorum ve tamamen katılıyorum" ifadelerinin sıralı tepkiler ifade ettiğinin de bir göstergesi olarak yorumlanabilir. KTM'ye göre ölçekleme sonucu, ölçekten elde edilen maksimum bilgi 48.23 ve bu bilgi miktarını veren tutum düzeyi -1.40'dır.

Yapılan çalışma, öğretmen adaylarının ölçme ve değerlendirme dersine yönelik tutumunu belirlemek üzere MTK'ya dayalı olarak geliştirilen 33 maddelik ÖDET'in oldukça geçerli ve güvenilir bir ölçme aracı olduğunu göstermektedir.

MTK'ya dayalı bir model olan KTM ile ölçeklenen ölçeğin model veri uyumunun sağlanması, ölçeğin farklı gruplarda uygulansa da değişmez parametre kestirimleri elde edilmesini sağlar. Bu araç, yeni haliyle farklı öğretmenlik programlarındaki aday öğretmenlerin ölçme ve değerlendirme dersi öncesi ve sonrası tutumlarının karşılaştırmasını konu alan araştırmalarda, olumsuz tutuma sahip olduğu saptanan gruplarda dersin öğretimi sürecinde yürütülecek çeşitli manipülasyonların etkisinin değerlendirileceği çalışmalarda geçerli ve güvenilir bir araç olarak kullanılabilir. Bunun yanında,

ölçme ve değerlendirme derslerindeki başarı ve bu derse yönelik tutum arasındaki ilişkinin inceleneceği araştırmalarda da kullanılabilir.