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:

$$X = T + E \tag{1}$$

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).

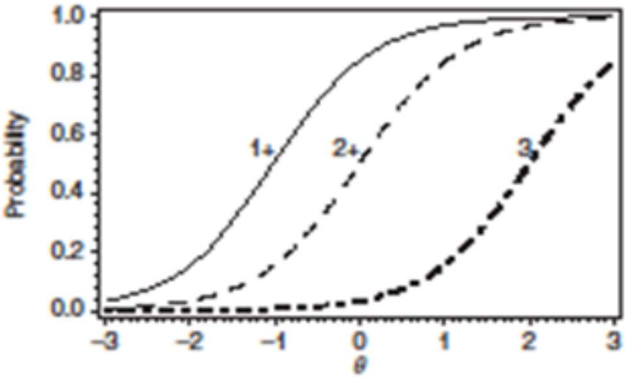

Figure 1. *GRM model for a 4-category item (ranked between 0-3). (Excerpt from DeMars, 2010).*

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.

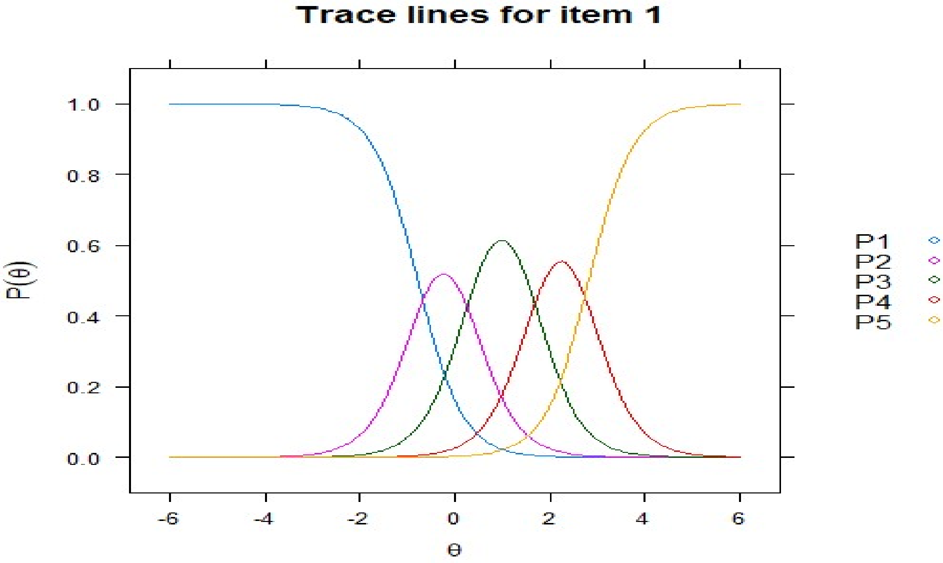

Figure 2. *All category/score possibilities for a 5-category item. These probabilities are calculated using item limit values or item threshold values.*

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.

$$\boldsymbol{P_{ik}^{*}(\theta)} = \frac{\boldsymbol{e^{1.7a_i(\theta-b_{ik})}}}{\boldsymbol{1+e^{1.7a_i(\theta-b_{ik})}}} \qquad (2)$$

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 $\mathrm{I_i}$ *(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).

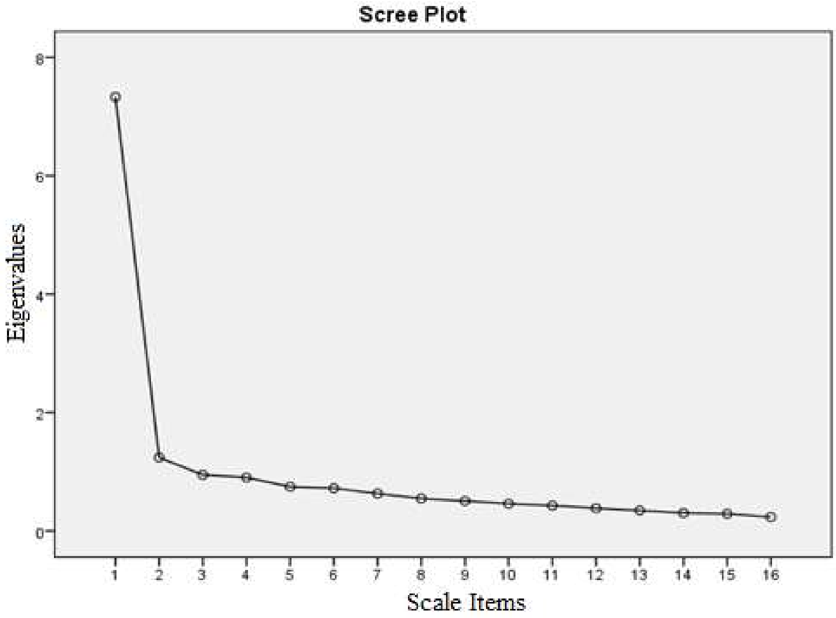

Figure 3. *Eigenvalue graphic*

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}	b_{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 $\theta = -0.780$, while the ability level to respond with a 50% probability in Category 5 is $\theta = 2.828$.

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.

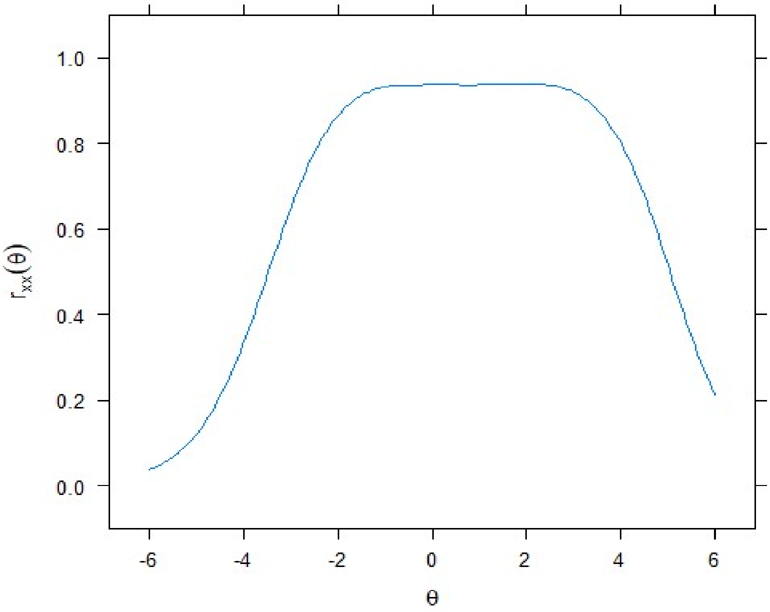

Figure 4. *Marginal reliability coefficient of PSS according to GRM*

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.

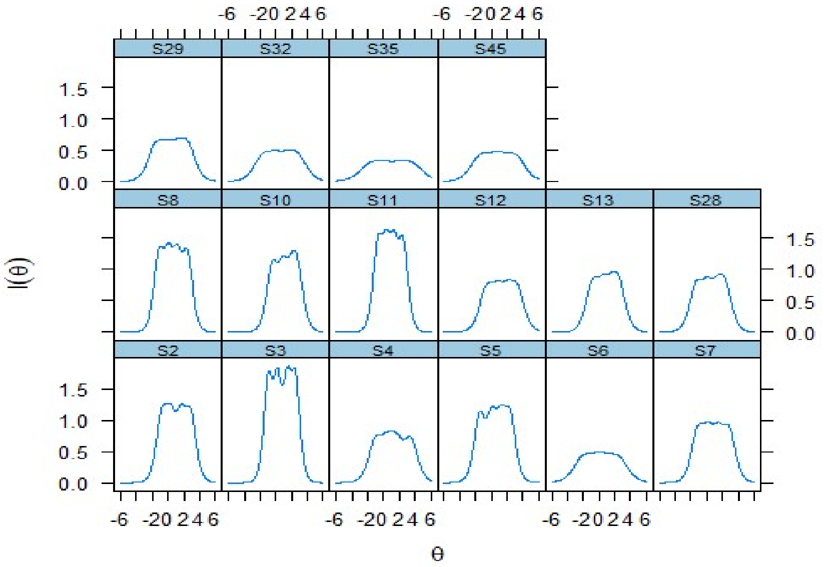

Figure 5. *PSS Item information functions*

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.

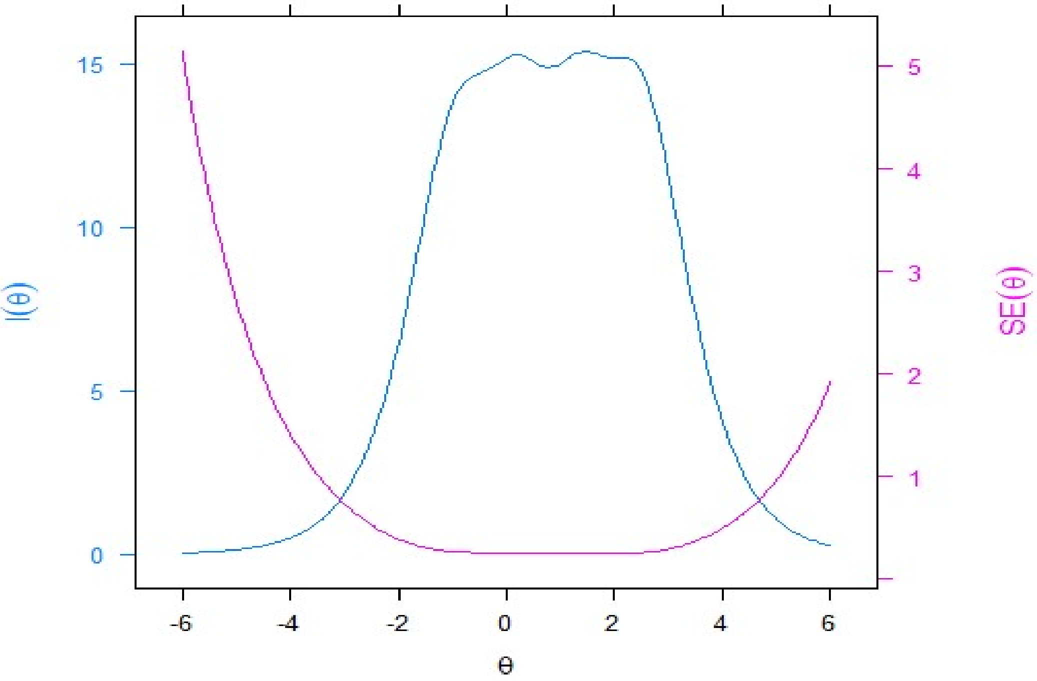

Figure 6. *Test information function*

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