ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 129-146

Investigation of Psychometric Properties of Likert Items with the Same Response Categories Using Polytomous Item Response Theory Models *

Esra SÖZER ** Nilüfer KAHRAMAN ***

Abstract
The purpose of this study was to investigate within- and between-threshold parameter invariance for items of a fourteen-item Positive Affect Scale developed to assess positive moods (like happy, peaceful, etc.) of university students. To test whether the estimated threshold parameters were as expected (1 to 5, with increments of 1) across all the 14 items, Graded Response, Partial Credit, and Rating Scale Models were fit the response data collected from 326 students. A comparison of the model fit statistics, such as the negative 2log likelihood and chi-square values, revealed that the Graded Response Model had the best fit and that the thresholds estimates for all the items in the Positive Affective Scale were reasonably close to the expected 1 to 5 values with increments of 1. The study illustrates how polytomous response models can be used to test the psychometric quality of items with ordinal rating scales.

Key Words: Item parameters, positive affect, polytomous, threshold, item response theory.

INTRODUCTION

When the response scales of the polytomous scored items are formulated, e.g., Likert scale, it is expected that respondents will choose the category that best describes their state given the measured trait. Even if it can be argued that this is a reasonable expectation, there remain several unanswered questions about how individuals' self-ratings compare amongst themselves, related to potential differences that may exist in the decision-making processes of the individuals when evaluating their state given the scale provided. The study of defining and testing for such individual differences has long been the focus of many scaling studies (e.g., Wang, Wilson, & Shih, 2006), all underlining the importance of a careful analysis of the scale properties of items, especially when subjective assessments are involved (Wang et al., 2006). Even when constructing ordinal scale assessment tools, the main objective of the psychometric work is about deriving the most accurate and meaningful information from the item responses (Wu & Adams, 2006).

Researchers studying traits from the affective domain do often face a greater number of challenges when evaluating the quality of their assessment results when compared to those who study traits from the cognitive domain, yet new methodological advancements rarely target their issues first. In this context, polytomous Item Response Theory (IRT) models, commonly used in calibrating items of most cognitive assessment tools, are yet to gain such common use when it comes to calibrating ordinal rating scale items, which are often used in the evaluation of psychological constructs, such as personality traits (Baker, Rounds, & Zevon, 2000). Given that assessment tools assessing psychological characteristics are, in general, composed of rating scale items, it would be most reasonable that polytomous IRT models are used in estimating non-linear relationships between the

* This study was presented in 6th International Congress on Measurement and Evaluation in Education and Psychology (5-8 September, 2018 in Prizren/Kosovo).
** Res. Assist. PhD., Bartin University, Faculty of Education, Bartin-Turkey, esrsozer@gmail.com, ORCID ID: 0000-0002-4672-5264
*** Prof. PhD., Gazi University, Gazi Education Faculty, Ankara-Turkey, kahramannilufer@gmail.com, ORCID ID: 0000-0003-2523-0155

To cite this article:
Sözer, E., & Kahraman, N. (2021). Investigation of psychometric properties of likert items with the same response categories using polytomous item response theory models. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 129-146. doi: 10.21031/epod.819927

Received: 02.11.2020
Accepted: 16.04.2021

propensity level of the respondent and the likelihood of responding in a certain category (Embretson & Reise, 2000).

The prototypical Likert-type scale has five categories. These are printed equally spaced and equally sized on the response form (Figure 1). The intention is to convey to the respondent that these categories are of equal importance and require equal attention (Linacre, 2002). Response categories have an explicit and clear continuum and reveal the underlying psychological structures of these categories.

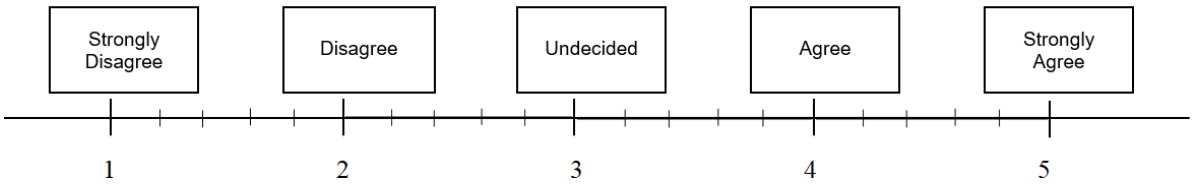


Figure1. Likert-Type Scale Response Categories

According to Linacre (2002), from a measurement perspective, the rating scale may appear in different forms (Figure 2). The rating categories still have a continuum and attempt to measure a psychological construct. Since the psychological construct intended to be measured conceptually is infinitely long, the two extreme categories are also infinitely wide. However, individuals are predominantly in the *agree* category. The size of intermediate categories such as *undecided* is dependent on how they are perceived and used by the respondents. *Agree* categories are usually more attractive than *disagree* categories. Therefore, *agree* categories may be represented by a wider interval for the measured psychological construct.

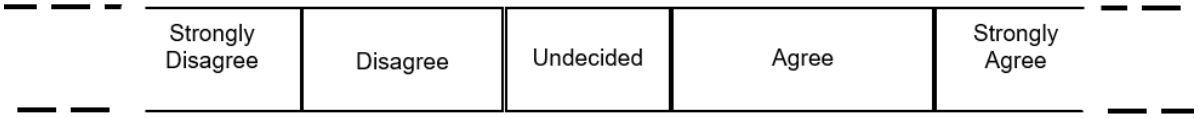


Figure 2. Typical Likert Scale Response Categories from Measurement Perspective

How the variable is divided into categories affects the reliability of a scale (Linacre, 2002). The rating categories with equal intervals as in Figure 1 or ordinal as in Figure 2 can be analyzed with polytomous IRT models. Polytomous IRT models are needed to represent the nonlinear relation between examinee trait level and the probability of responding in a particular category (Embretson & Reise, 2000). Polytomous models allow the use of different item discrimination values in weighting items, the estimation of measurement errors at each ability level, and achieving parameter invariance for the individuals and items (Lord, 1980).

Polytomous models vary based on whether the response categories are ordinal or non-ordered. In this case, in each model, the meaning of the response probability obtained for the response categories will also differ within the context of parameters that the model allows defining. The Graded Response Model (GRM; Samejima, 1969), one of the polytomous models used for modelling ordered response categories, the likelihood of marking each category or an upper category is modelled; while in Partial Credit Model (PCM; Embretson & Reise, 2000), the likelihood of scoring or choosing each category is directly modelled (instead of the category or an upper category).

In this study, category threshold parameters between consecutive categories estimated according to the GRM model used in the estimation of scale item parameters represent the ability level required for responding to the category and above with a probability of .50. According to PCM, the items are assumed to have equal discrimination (slope). In this case, the probability of an individual's responding to a category is computed as a function of the difference between an individual's ability level and the

category threshold parameter (step difficulty). Unlike GRM, step difficulty parameters represent the relative difficulty of each step. According to Rating Scale Model (RSM), the last model used in the study, the location parameter estimated separately for each item reflects the relative easiness or difficulty of the particular item. In this model, it is assumed that the same response format is used for all items in the scale; therefore, category threshold values are estimated on an equal basis for all items. In RSM, the response likelihood of an item is determined by location parameter and category threshold parameter (Embretson & Reise, 2000).

Item response categories with different properties are analyzed with different measurement models mentioned above, and model-data fit is assessed. In addition to the assessment of a model-data fit, it is emphasized that the importance of including basic observations to determine to what extent the model fits the psychological reality that underlies the responses (i.e., response format) (Samejima, 1996). For this reason, it is important to determine the characteristics of the analyzed item response categories (whether the categories have a similar order for each item) and to what extent they fit the psychological structure they are trying to measure, in terms of the reliability and validity of the measurement results obtained.

A review of the literature showed that polytomous IRT models are widely used in analyzing psychometric properties of Likert-type rating scales (de Ayala, Dodd, & Koch, 1990; Koch, 1983). These models are also used for analyzing psychometric properties of measurement tools designed for measuring affective skills such as self-esteem (Gray-Little, Williams, & Hancock, 1997), emotional regulation (Rubio, Aguado, Hontangas, & Hernandez, 2007), self-identification (Flannery, Reise, & Widaman, 1995), emotional intelligence (Cho, Drasgow, & Cao, 2015), subjective well-being (Baker et al., 2000), self-reflection (Silvia, 2021), anxiety (Caycho-Rodríguez et al., 2021) as well as of those for measuring cognitive skills (Min & Aryadoust, 2021). Few studies were found in our country which employed polytomous IRT models for analyzing psychometric properties of measurement instruments used for emotional skills. It was found that polytomous IRT models were used for developing and adapting measurement tools like resilience scale (Yaşar & Aybek, 2019), attitude scale (Demirtaşlı, Yalçın, & Ayan, 2016); however, the properties of item response categories were not analyzed in many scale development and adaptation studies. This study focused on the importance of this issue and elucidated how the studies could be conducted in practice by exemplifying through a scale in the context of the use of polytomous IRT models in measuring constructs related to the affective domain such as subjective well-being.

Positive Affect Scale (PAS) used in this study is designed similarly to the Positive and Negative Affect Scale (PANAS; Watson, Clark, & Tellegen, 1988), but it is a five-point graded (1-5, with increments of 1) Likert scale consisting of 14 positive affect items. These self-report constructs by which individuals assess themselves are considered substantial individual differences' variables for a long time (Hattie, 1992). Determination and improvement of positive affects of individuals such as subjective well-being, happiness, and resilience are among the main objectives of education environments. The responses to polytomous scoring items used for analyzing affective characteristics are based on subjective assessments by which individuals are assumed to select the categories which describe them best. At this point, the satisfaction of the assumption that the order between response categories in the scale used is the same for each item (e.g. evenness of threshold parameters between 1 and 2, 2 and 3, ...) and that the order between items refers to the same meaning is important for a reliable interpretation of measurement results (Koch, 1983).

Purpose of the Study

The purpose of this study was to investigate response categories of rating items (from 1 to 5) in a 14-item PAS scale developed to measure positive affects and to demonstrate the extent of similarities/differences between these categories regarding the items. It was aimed to obtain an estimation of item parameters for polytomous scoring items in PAS scale utilizing different polytomous models, analyze model-data fit and make a comparative evaluation of the measurement precision at different ability levels across the affect scale. Considering the polytomous response format

of PAS and theoretical relationship between polytomous models and response processes, whether category threshold parameters used for determining responses to the items were ordered in inter-item was tested through GRM (Samejima, 1969), PCM (Embretson & Reise, 2000) and RSM (Andrich, 1978). Based on the requirements set out by each of these models, the validity of the assumption of invariance of category threshold parameters for all items was analyzed using the data in practice.

METHOD

This study is designed as a descriptive comparative study that analyzed psychometric properties of the PAS according to polytomous Item Response Theory models (Glass & Hopkins, 1984; Kaptan, 1995).

Study Group

The study group comprised 326 volunteer students (pre-service teachers) who studied at the Gazi Faculty of Education in the academic year 2017-2018. The study group included 166 female (51%) and 52 male (17%). The participants were in an age range of 19-35 years. Among these participants, 6 of them were 19 years old (1.8%), 77 were 20 years (23.6%), 92 were 21 years (28.2%), 24 were 22 years (24%), 7 were 23 years (2.1%), 3 were 24 (0.9%), 2 were 25 years (0.6%), 1 participant was 28 years (0.3%), 3 participants were 29 years (0.9%), 2 were 30 years (0.6%) and 1 participant was 35 years (0.3%) old. (Demographic information about the study group was obtained by a separate scale and was not mandatory. Therefore, the values for those whose information could be reached were presented.)

Data Collection Tools

The data used in this study come from a more comprehensive study called Emotion Ruler Field Study (Kahraman, Akbaş, & Sözer, 2019). Positive and Negative Affect Scale consists of 27 positive and negative affects. The individuals were asked to mark the best describe them among the response categories (from 1 for *very slightly or not at all* to 5 for *extremely*). According to the results of Exploratory Factor Analysis (EFA) for factor structure of the scale, a Kaiser-Meyer-Olkin (KMO) value was found to be 0.94. Chi-square ($\chi 2$) statistic and the result of Bartlett's test was statistically significant (χ^2 (351) = 5605.97, $p < .05$). The data were found to have a two-factor structure with eigenvalues of 11.13 and 3.53. The total variance explained by the factors was 51%. Confirmatory Factor Analysis (CFA) results used for verifying factor structure showed that model-data fit was at an acceptable level, and the scale had a two-factor structure (χ^2 (294) = 838.76, RMSEA= .08, CFI = .87, TLI = .86 and SRMR = .08). The results of Cronbach's Alpha correlation coefficients showed that the reliability for each factor was respectively for positive and negative affects .92 and .91.

In this study, the data came from positive affect items was employed. This sub-factor named PAS consists of 14 items that ask individuals to mark one of the response categories (from 1 for *very slightly or not at all* to 5 for *extremely*) for each item given to them. 14 positive affects included in the scale are as follows (Table 1): Happy, peaceful, contented, open to communication, understanding, motivated, resilience, strong, self-confident, determined, successful, optimistic, brave and energetic. Descriptive statistics for items are given in Table 1. Analyses for the factor structure of PAS are presented in the data analysis section.

Data Collection Procedure

Data for the PAS were collected from the participants through an online application. PAS consists of self-report items whereby individuals are asked to choose one of the response categories appropriate for them.

Table 1. Descriptive Statistics for Positive Affect Scale

Items	Mean	S.D.	Skewness	Kurtosis	r_{ij}*
1. Happy	3.17	0.95	-.34	-.04	.68
2. Peaceful	3.03	1.08	-.25	-.67	.63
3. Contented	2.98	1.06	-.21	-.58	.68
4. Open to communication	3.60	0.99	-.50	-.18	.58
5. Understanding	3.57	0.91	-.41	-.07	.51
6. Motivated	3.10	1.05	-.08	-.49	.74
7. Resilience	3.52	0.99	-.42	-.27	.68
8. Strong	3.48	1.05	-.44	-.41	.66
9. Self-confident	3.31	1.05	-.23	-.42	.66
10. Determined	3.27	1.09	-.28	-.49	.66
11. Successful	3.22	1.00	-.23	-.13	.60
12. Optimistic	3.38	1.03	-.24	.-.54	.60
13. Brave	3.20	1.06	-.12	.-.54	.63
14. Energetic	2.72	1.09	.17	-.62	.62

* r_{ij} = correlation values for item-total test score

Data Analysis

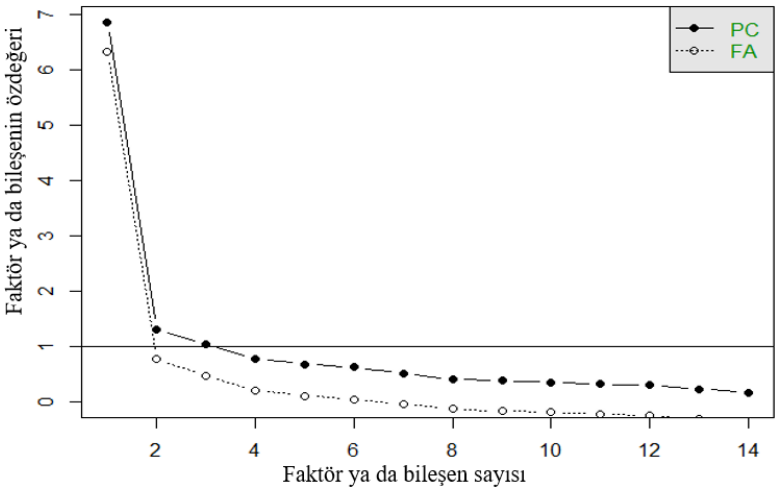
The data were analyzed using the "mirt" package (Chalmers, 2012) in the R (R Core Team, 2016) program. Item parameters for PAS were estimated using GRM (Samejima, 1969), PCM (Embretson & Reise, 2000) and RSM (Andrich, 1978). Descriptive statistics (mean, standard deviation) obtained at the initial data analysis stage, and correlation values for item-total test score (r_{ij}) are given in Table 1. Besides, the factor structure of the scale (unidimensionality assumption) was analyzed using EFA, CFA and parallel analysis. The reliability coefficient for PAS was determined as a Cronbach's α value of .92. In the evaluation of model-data fit for factor analysis, RMSEA ≤ .08 (Steiger & Lind, 1980); SRMR ≤ .08 (Brown, 2015); CFI ≥ .90 (Hu & Bentler, 1999) and TLI ≥ .90 criteria were considered.

An examination of descriptive statistics given in Table 1 shows that skewness and kurtosis coefficients are in the range of ±1. This points out a normally distribution of the data. In the second stage, IRT models used in parameter estimation and model-data fit process are presented.

Unidimensionality assumption

Unidimensionality which is the fundamental assumption of unidimensional IRT models was analyzed using EFA, CFA and parallel analysis. The KMO value was found to be 0.91, and according to Bartlett's test result, χ^2 value was significant (χ^2 (91) = 2642,29, $p < .05$). The dimensionality of data structure was examined using a scree plot (Figure 3), and a single-factor structure with an eigenvalue of 6.85 was identified. Total variance explained by the factor was 49%, and factor loadings for the items varied between .53 and .77.

Scree plot indicates a rapid decrease in the eigenvalue from the first to the second factor. This shows that PAS had a dominant single-factor structure. At the end of CFA performed to verify factor structure, it was confirmed that model-data fit was at an acceptable level and the scale had a single-factor structure (χ^2 (74) =283.79, RMSEA = .08; CFI = .91, TLI = .88 and SRMR = .06).

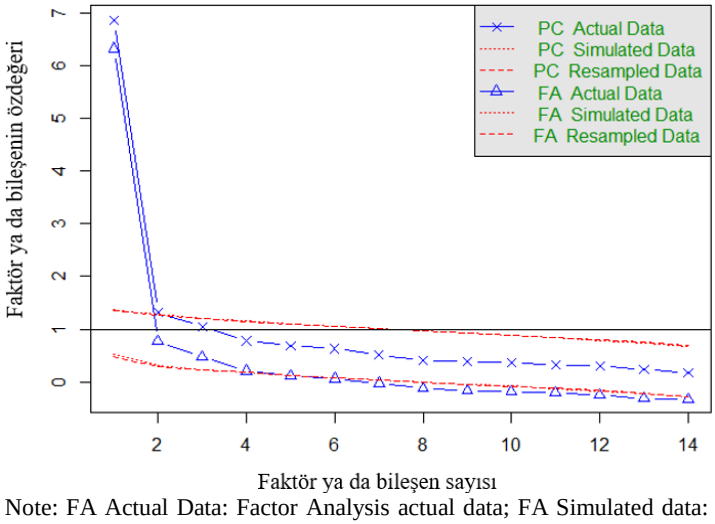


Note. FA: Factor Analysis; PC: Principal Component Analysis

Figure 3. Scree Plot of the PAS Factor Structure

Parallel analysis

Parallel analysis generates random correlation matrices and conducts a factor analysis with these matrices followed by a comparison of eigenvalues obtained through observation of real data with those obtained from simulated data. The fact that eigenvalues obtained from real data are higher than simulated data signals the existence of significant factors.



Note: FA Actual Data: Factor Analysis actual data; FA Simulated data: Factor Analysis simulated data; PC Actual Data: Principal Component Analysis actual data; PC Simulated Data: Principal Component Analysis simulated data

Figure 4. Parallel Analysis Scree Plot

Red-dotted lines in Figure 4 indicate values for simulated data, and blue-dotted lines indicate values for actual data. Blue dots derived from factor analysis up to the red line for simulated data (triangular shape) show factors and components obtained from the data. As a result of the analysis, it was concluded that a single-factor structure was provided.

Local independence assumption

Local independence, given a constant ability level that affects test performance, means that individuals' responses to items are independent of each other. Local independence often occurs when an item is an answer to another item or items depend on a scenario or reading text (DeMars, 2010). Various statistics such as Yen's Q_3 (1984) are suggested for analyzing local independence assumption. The Q_3 statistic proposed by Yen takes into account the relationships between item pairs. First of all, parameters for items and individuals are estimated through an IRT model that is fit for the data. After the estimation of parameters, a residual matrix is formed using the residuals of each item, and correlations between them can be analyzed (DeMars, 2010). If the local independence assumption is confirmed, the items will be independent of each other given an ability level (θ) condition.

It is stated by various studies that if the unidimensionality assumption is met, the local independence assumption is also met (Embretson and Reise, 2000; Hambleton & Swaminathan, 1985). At this point, it was verified by the results of factor analysis that items used in the study displayed a unidimensional structure. Since the unidimensionality assumption was met, it was assumed that the local independence assumption was also met.

Parameter estimation

In the second stage of the analysis, psychometric properties of response categories of 14 items were analyzed using GRM, PCM and RSM. Brief information about the models used in the analysis is given below.

Graded response model (GRM): GRM was used firstly for the estimation of item and test parameters. GRM is appropriate to use when item responses can be characterized as ordered categorical responses. The best advantage of GRM lies in that it provides more information about the ability of individuals compared to dichotomous models. Polytomous items are categorically similar to dichotomous items, but they have more than two response categories. These ordered categories have a *k*-1 boundary or threshold parameters that separate the categories for an item with k ordered response categories. In comparison with the probability of an individual to respond to any categories lower than a certain category level, they attempt to determine the likelihood to respond to that category or to those above that category (DeMars, 2010).

In the GRM, each scale item (*i*) is described by two parameters. First, the a_i (discrimination) parameter can be defined as the variation strength of response probability as a function of the latent trait (Rubio et al., 2007). Second, b_i (threshold parameter) refers to the level of latent trait, θ, at which, for each category boundary, the probability of giving a positive response rather than a negative one to that boundary is .5 (Embretson & Reise, 2000).

GRM requires a two-stage procedure to computing the category response probabilities (Embretson & Reise, 2000). In the first step, the estimation of response probabilities involves the computation of *k*-1 curves for each item of the form given in Equation 1.

$$P_{ik}^{*}(\theta_j) = \frac{e^{Da_i(\theta_j - b_{ik})}}{1+e^{Da_i(\theta_j - b_{ik})}} \quad (1)$$

b_{ik} parameter, for each category boundary, is the level of the latent trait, θ, at which the probability of giving a positive response rather than a negative one to that boundary is .5. $P_{ik}^{*}(\theta_j)$ (operating characteristic curve) refers to the probability of an individual with θ_j to respond above a determined k category boundary. In Equation 2, category characteristic curves are estimated in the second stage, and they represent the probability of an examinee responding in a particular category conditional on trait level. $P_{ik}(\theta_j)$ refers to the probability of an individual under θ_j condition to choose a k category of item *i* (Embretson & Reise, 2000).

$$P_{ik}(\theta_j) = P_{ik}^{*}(\theta_j) - P_{i(k+1)}^{*}(\theta_j) \quad (2)$$

In this study, the Marginal Maximum Likelihood (MML) method was used for the estimation of GRM item parameters. In GRM, discrimination (slope) for each item and 4 threshold parameters for 5-point response categories were estimated. It is assumed that inter-category threshold (*b*) parameters for each item are ordered in GRM (Embretson & Reise, 2000).

Partial credit model (PCM): PCM was used secondly in the estimation of item and test parameters. PCM (Muraki, 1992) was 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 personality traits (Embretson & Reise, 2000).

It is an extension of the Rasch Model, and raw scores are sufficient for the estimation of ability levels. In this model, the individuals with the same raw scores are at the same ability level. Unlike GRM, the discrimination (a_i) parameter is assumed to be equal for all items. The likelihood of responding to a category can be directly modelled. PCM is a divided-by-total or, as we term it, a direct IRT model (Embretson & Reise, 2000). This means that the probability of responding in a particular category will be written directly as an exponential divided by the sum of exponentials. Assume that item *i* is scored $x = 0 \ldots m_i$ for an item with $K_i = m_i+1$ response categories. For $x = j$ the category response curves for the PCM can be written as in Equation 3.

$$P_{ix}(\theta) = \frac{\exp \sum_{j=0}^{x}(\theta-\delta_{ij})}{\sum_{r=0}^{mi} \exp \sum_{j=0}^{r}(\theta-\delta_{ij})} \quad (3)$$

In PCM, different from GRM, step difficulty is defined instead of category threshold parameter. In Equation 3, $\sum_{j=0}^{0}(\theta - \delta_{ij}) = 0$ terms are called the item step difficulty associated with a category score of j. Step difficulty can be directly interpreted as the point on the latent trait scale at which two consecutive category response curves intersect. Step difficulty can also be defined as the difficulty parameter for passing from one category to the other (Embretson & Reise, 2000).

MML method was also used in PCM for the estimation of item parameters. In PCM, since the discrimination (slope) parameter is considered equal for all items, one discrimination parameter is estimated for all items. *k*-1 step difficulty (b) estimation is obtained for an item with k ordered response categories.

Rating scale model (RSM): It can be used when the items in the scale have the same response format (Embretson & Reise, 2000). In this model, step difficulties of the PCM are defined by location parameter that indicates the place of the item on ability scale and category threshold parameter between consecutive categories. Each item has a single scale location parameter which reflects the difficulty or easiness of the particular item. By the way, the scale location parameter indicates the distance of averages of step difficulties across consecutive categories to zero. It is equivalent to a limited version of PCM where category threshold parameters are equal across items. As is the case in PCM, item discrimination (ai) parameters do not vary across items.

In RSM, the item discrimination parameter is considered equal for all items. k-1 category threshold parameters (*b*) estimation is obtained for an item with *k* ordered response categories. Since the same scale format is used for all items, category threshold parameters are assumed to be equal for all items. Step difficulty, on the other hand, is defined as the sum of item-specific location parameters and category threshold parameters. MML method was also used in RSM for the estimation of item parameters.

In the RSM model, the step difficulties of the PCM are decomposed into two components, namely, l_i and d_j, where $d_{ij} = (l_i + d_j)$. The l_i is the location of the item on the latent scale and the d_j are the category threshold parameters (Embretson & Reise, 2000). RSM is written as Equation 4.

$$P_x(\theta) = \frac{\exp\left\{\sum_{j=0}^{x}[\theta-(\lambda_i+\delta_i)]\right\}}{\sum_{x=0}^{M} \exp\left\{\sum_{j=0}^{x}[\theta-(\lambda_i+\delta_i)]\right\}} \quad (4)$$

In PAS with ordered and 5-point Likert type response categories, the same response categories (also in the same number) are used for all scale items. Therefore, item and test parameters were analyzed

using GRM, PCM and RSM in an attempt to determine the best fit model to be used for analyzing psychometric properties of the scale.

Model - data fit

For assessment of model-data fit, -2loglikelihood values of polytomous model pairs were compared. Firstly, a comparison was made based on GRM and RSM -2loglikelihood values, χ^2 value and degrees of freedom. AIC and BIC values were also examined. Subsequently, GRM and PCM models were compared. Also, standard error and parameter invariance was investigated. For measurement precision, the amount of information provided by each item across different ability levels was evaluated along with item information functions. The ordinal state of item response categories for each item was examined employing graphical methods.

RESULTS

14 items in PAS were scaled using three different polytomous IRT models. Table 2 displays the model-data fit statistics and Table 3 displays the amount of item information for each model.

Model-data fit was evaluated by comparing in model pairs of lower AIC, BIC and -2loglikelihood values from the models. According to AIC and BIC values in Table 2, the models with the lowest AIC values are GRM, RSM and PCM, respectively, while the models with the lowest BIC values are RSM, GRM and PCM, respectively. These results show that GRM and RSM fitted the data better than PCM.

Table 2. Model-Data Fit Indexes for Polytomous IRT Models

Models	AIC	BIC	χ^2	Degrees of freedom (*df*)
GRM	**11454.84**	11722.66	5763.32	70
RSM	11562.64	**11631.51**	5657.42	19
PCM	11575.21	11793.30	5730.61	57

Table 3 presents item and total test information amount and marginal reliability values derived from different models. The highest amount of total test information was obtained from RSM. Other information amounts were provided by GRM and PCM, respectively. Also, although the reliability coefficient of all three models was close to each other, the highest reliability coefficient was obtained with GRM with a value of .93. Firstly, the values obtained from RSM and GRM which provided the highest amount of total test information were compared to -2loglikelihood, degrees of freedom (*df*) and χ^2 values. The number of parameters varies depending on the different models.

Table 3. Amount of Item and Total Test Information from Polytomous IRT Models

Items	**GRM***	**PCM****	**RSM*****
1. Happy	7.20	3.99	3.99
2. Peaceful	5.44	3.99	4.82
3. Contented	7.08	3.99	5.67
4. Open to communication	4.49	3.99	28.95
5. Understanding	3.70	3.99	22.20
6. Motivated	8.79	3.99	4.15
7. Resilient	6.90	3.99	15.56
8. Strong	6.33	3.99	11.52
9. Self-confident	5.73	3.99	5.08
10. Determined	5.57	3.99	4.53
11. Successful	4.76	3.99	4.14
12. Optimistic	4.64	3.99	6.81
13. Brave	4.89	3.99	4.05
14. Energetic	4.79	3.99	23.54
Total Information	80.35	55.98	145.08
Marginal Reliability	.93	.92	.92

* GRM: Graded Response Model; **PCM: Partial Credit Model; ***RSM: Rating Scale Model

According to RSM, a common a_i parameter, (the number of categories (5) - 1 = 4) category threshold parameters and location parameters for each item were estimated, and the degrees of freedom is (19). In GRM, the a_i parameter for each item and (the number of categories (5) – 1 = 4) category threshold parameters for each item were estimated, and the degrees of freedom was determined as (70). According to this, χ^2 (70, 19) = 5763.32 - 5657.42 = 105.9 and approximate table χ^2 value, χ^2 (51, .05) = 67.50. The difference between the -2loglikelihood χ^2 values from model pairs was found to be significant. Therefore, it can be concluded that GRM is more appropriate for the data.

Secondly, the difference in -2loglikelihood χ^2 values obtained from GRM and PCM was compared with χ^2 statistic using the .05 significance level and degrees of freedom. While the degrees of freedom was determined as (70) for GRM; in PCM, a common a_i parameter for each item and (the number of categories (5) – 1 = 4) category threshold parameters were derived for each item, and the degrees of freedom was determined as (57). In this case, χ^2 (70, 57) = 5657.42 – 5730.61 = -73.19 and, approximate table χ^2 value, χ^2 (13; .05) = 22.36. The difference between the -2loglikelihood χ^2 from model pairs is not significant. This indicates that there is no difference between GRM and PCM. Furthermore, in GRM, the reliability and maximum information values were found to be .93 and 80.35, respectively with a lower AIC value. As a result of model pair comparisons, it was determined that GRM fits the data better, and parameter estimations were performed using GRM. Using GRM, a_i parameter (discrimination) for each item and 4 threshold parameters for 5-point response categories were estimated. Table 4 shows estimated parameters for PAS items.

In GRM calibration, 70 parameters were estimated. Item discrimination parameter refers to the item's power of sorting individuals based on their abilities across latent trait scale. 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 above 1.70 (Baker, 2001). Item discrimination (a_i) parameters for 14 items vary between 1.25 and 2.66 and with item 6 having the highest and item 5 having the lowest level. Accordingly, it is understood that discrimination values of items are of medium and high levels. In the context of data structure, the a_i parameter can be considered as the numerical value of the psychological uncertainty of an item (Roskam, 1985). Higher a_i parameter values indicate that the item has a well-defined and clear meaning (Ferrando, Lorenzo, & Molina, 2001). As a result, it was concluded that 14 items in the scale were well-defined items with high discrimination.

Table 4. Estimated Item Discrimination and Category Threshold Parameters According to GRM

Items	a_i (se)	b_1(se)	b_2(se)	b_3(se)	b_4(se)
1. Happy	2.21(.20)	-1.94(.15)	-1.06(.18)	0.38(.12)	2.03(.44)
2. Peaceful	1.83(.17)	-1.75(.15)	-0.70(.15)	0.40(.11)	2.14(.34)
3. Contented	2.25(.21)	-1.58(.12)	-0.68(.14)	0.48(.11)	2.00(.22)
4. Open to communication	1.55(.16)	-2.92(.30)	-1.57(.29)	-0.30(.23)	1.36(.40)
5. Understanding	1.25(.14)	-3.71(.46)	-1.97(.40)	-0.26(.32)	1.82(.62)
6. Motivated	2.66(.24)	-1.77(.12)	-0.76(.16)	0.40(.11)	1.56(.03)
7. Resilient	2.18(.20)	-2.42(.19)	-1.30(.23)	0.13(.18)	1.27(.50)
8. Strong	2.08(.19)	-2.27(.18)	-1.14(.21)	-0.12(.16)	1.27(.47)
9. Self-confident	1.92(.18)	-2.18(.17)	-1.07(.20)	0.19(.14)	1.44(.68)
10. Determined	1.92(.18)	-1.98(.16)	-0.99(.18)	0.19(.13)	1.48(.71)
11. Successful	1.64(.16)	-2.23(.20)	-1.21(.21)	0.41(.14)	1.83(.20)
12. Optimistic	1.58(.15)	-2.70(.26)	-1.21(.24)	0.10(.19)	1.56(.59)
13. Brave	1.68(.16)	-2.25(.19)	-0.94(.19)	0.39(.15)	1.70(.89)
14. Energetic	1.67(.16)	-1.56(.13)	-0.30(.11)	0.94(.21)	2.26(.92)

Note: a_i = item discrimination; se = standard error; b_i = category threshold

b_{ik} parameters (b_{i1} and b_{i4}) show the position of items in the latent trait (ability) scale. For example, for item 1, b_{11} = -1.94 refers to the ability level required to respond to category 1 and above with a likelihood of .50. b_{15} = 2.03 refers to the ability level required to respond to category 5 with a likelihood of .50. It is seen that along the latent trait scale, first category threshold parameter values were distributed around -2, second category threshold parameter values around -1, third category threshold parameter values around 0, and fourth category threshold parameter values were distributed around

1.5. This indicates that the scale better differentiates people across with the latent trait scale. Also, category threshold parameter values displayed a hierarchical increase along the ability scale. According to the results, it is understood that it is suitable to use GRM for measuring the psychometric properties of PAS.

Figure 5 presents category threshold parameters estimated for 14 items. a_i (discrimination) parameters obtained in GRM are treated as random effects. Since each item has its discrimination parameter value, graph lines belonging to the category threshold are not parallel to each other. However, it is seen in Figure 5 and Figure 6 that category threshold parameters of 14 items are ordinal for each item.

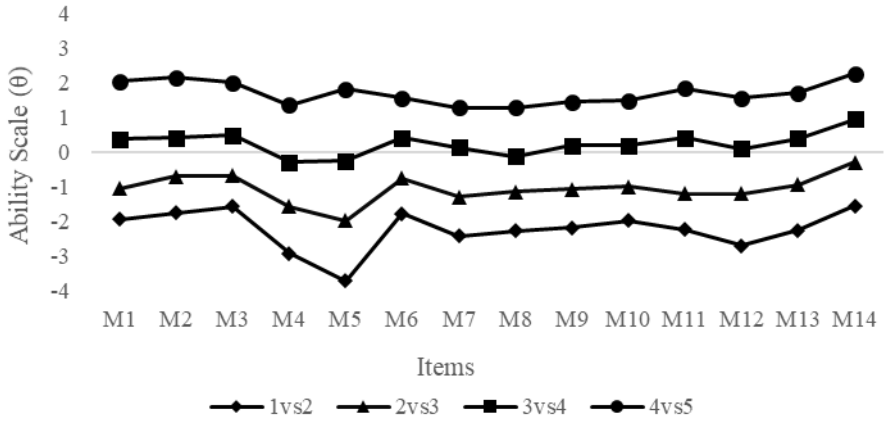


Figure 5. Order of Category Threshold Values for 14 Items Estimated by GRM

Figure 5, horizontal axis denotes 14 items and the vertical axis denotes ability (θ) scale. It is apparent in Figure 5 that category threshold parameters for the items of PAS are in a hierarchical order. In Figure 6, it is exemplified through item 2 and item 6 given that category threshold parameters are ordered based on item.

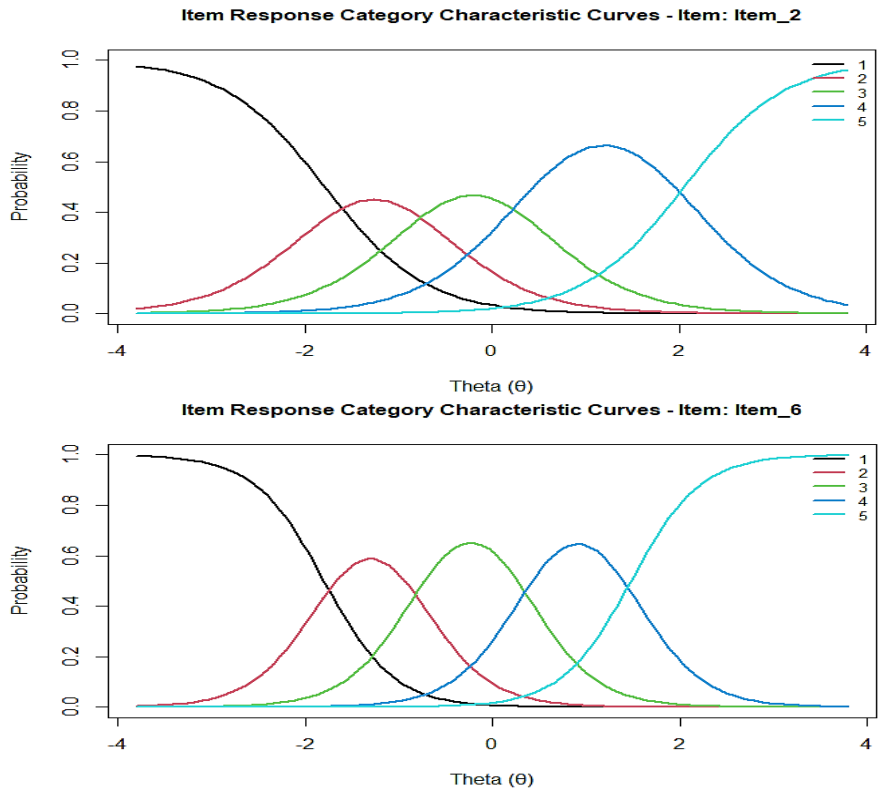


Figure 6. Item Response Category Characteristic Curves for Item 2 and Item 6

In Figure 7, item information functions are given for three items with high (Item 6), medium (Item 2) and low discrimination (Item 5) level. Figure 7 indicates how different discrimination (slope) values affect measurement precision throughout the ability scale. Accordingly, Item 6 with a high discrimination value provided more information than Item 2 and Item 5 all along the scale.

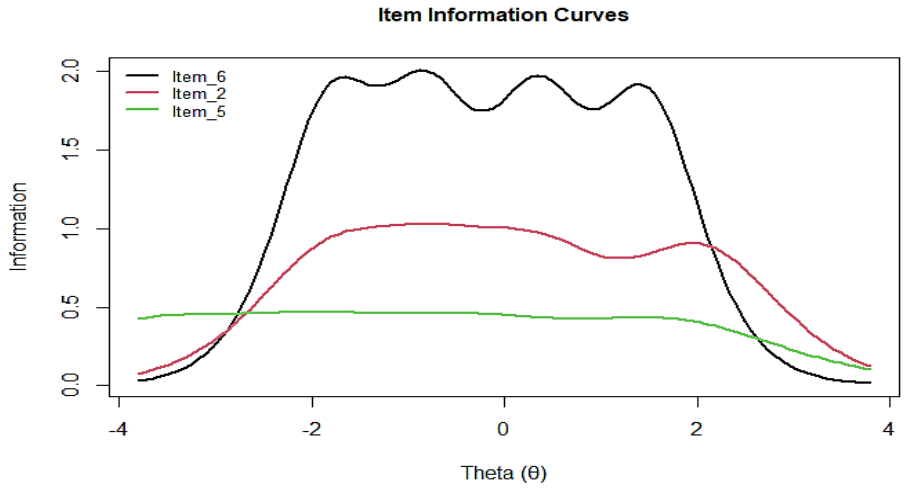
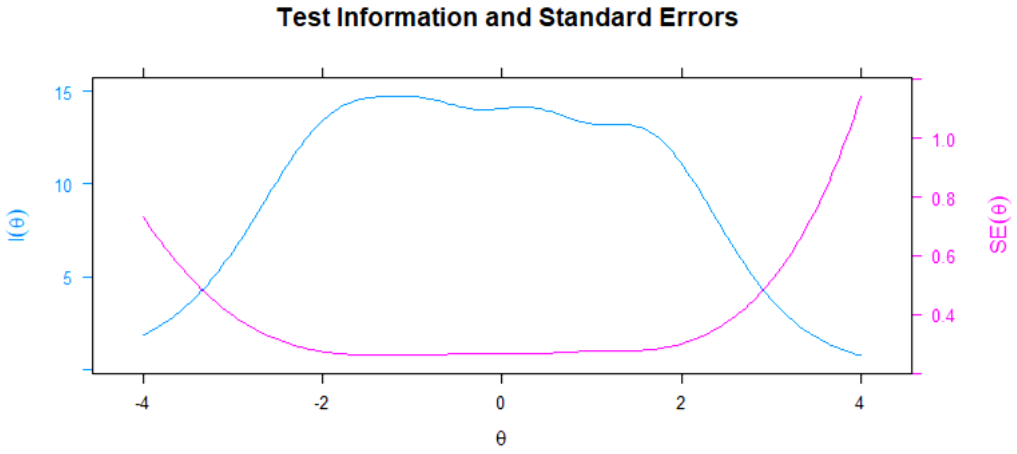


Figure 7. Item Information Curves with Low (Item 5), Medium (Item 2) and High (Item 6) Information Level

Figure 8 shows the relationship between the total test information of PAS based on GRM and the standard error. The amount of information obtained through the ability scale seems to be higher at the ability level within the interval of ($-2 \leq \theta \leq +2$). The figure also shows that standard error estimation is also lower in this ability level interval. It indicates that the amount of maximum information is provided by the scale around the ability level $\theta = (-1.40)$.



Note: θ = Latent trait scale (ability scale), blue line indicates total test information function (I(θ)), and the pink line indicates standard error (SE(θ)).

Figure 8. Test Information and Standard Errors for PAS Based on GRM

The sample was randomly divided into two groups to test parameter invariance and, then item discrimination and category threshold parameter values were estimated for each sub-group. Correlation between item discrimination values (a_i) from the two sub-groups is $r = .81$ ($p < .01$). Correlations between category threshold (b_{ik}) values were found to be $b_{i1} = 0.90$, $b_{i2} = 0.96$, $b_{i3} = 0.97$,

b_{i4}= 0.83 (p < .01), respectively. The results showed that the correlation values for parameters estimated from different samples were high; in other words, they were analogous, proving that parameter invariance was ensured.

DISCUSSION and CONCLUSION

The review of the literature on scaling reveals that there are many studies of cognitive test structures under IRT models. However, it is a fact that use of IRT-based models in developing scales for measurement of affective traits is relatively limited in our country (Demirtaşlı et al., 2016). The purpose of this study was to investigate whether category threshold parameters, which are used to determine responses to Likert-type polytomous items in measurement tools used particularly for measuring affective traits, were ordered within the items. Responses to polytomous items in Likert-type measurement tools assume that individuals choose the categories which best describe their states. However, differences may occur between assessments as individuals use different decision-making processes when making such decisions. It is important to employ appropriate methods and techniques for developing measurement tools to catch up with this variance between subjective assessments (Wang et al., 2006). The extent to which a psychological construct intended to be measured is represented by response categories of a measurement tool is very important in terms of psychometric properties. This study aimed to test the psychometric properties of the Positive Affect Scale used to determine positive affects across item response categories. The fact that item response categories in the scale are ordered for each item and have similar meanings is of importance for using and interpreting the results of the scale (Messick, 1995).

The ability levels required to respond to each category of each item are estimated separately for measurement tools scaled with IRT models. This allows achieving more reliable and valid results for the measurement of individual differences. The extent of fitness of response format in a measurement tool for the psychological reality which it intends to measure also affects the validity of measurements (Baker et al., 2000). Therefore, selecting the suitable model for the data is important for the interpretability of the inferences from the results. In this study, Samejima's GRM, PCM, and RSM were used for analyzing psychometric properties of item response categories. Results from different IRT models for scaling provide various information about categories. Psychometric properties of item response categories of Likert-type scale items within the scope of this study were evaluated to model-data fit within the context of specific parameters of each model. In particular, the analysis of inter-category psychometric properties of polytomous items used for measuring affective traits will also contribute to the significance of inferences from measurement results. Results based on different models which ensured model-data fit provide different information about the properties of categories.

Application data were used in this study, and the comparability of item parameters of 14-item PAS subject to the application was analyzed using polytomous IRT models. Model comparisons were made to determine the best fit IRT model for PAS items. As a result of analyses, GRM had to the best fit. Since the maximum amount of information provided by GRM and reliability of GRM is higher and its AIC value is lower, parameter estimations were made according to GRM in the analysis of psychometric properties. Similar results were obtained in various studies which examined psychological properties. In the study by Rubio et al. (2007), results that correspond to those of GRM were obtained in the analysis of psychometric properties of emotional adaptation scale, Rosenberg self-esteem scale (Gray-Little et al., 1997). GRM has been frequently used in the analysis of psychometric properties of measurement tools applied for analyzing response categories for positive and negative affects (Baker et al., 2000) and various affective traits (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Demirtaşlı et al., 2016; Köse, 2015).

Item discrimination parameters (a_i) for 14 items estimated based on GRM varied between the values of 1.25 and 2.66. Accordingly, the items had discrimination values of medium and high level. In the analysis, 4 category threshold parameters were estimated for each item. It is seen that along the ability scale, first category threshold parameter values were distributed around -2, second category threshold parameter values around -1, third category threshold parameter values around 0 and fourth category

threshold parameter values around 1.5. This shows that the scale well-distinguished people at different ability levels along the latent trait scale.

The information from test and item information functions proved to be higher at the ability levels in ($-2 \leq \theta \leq +2$) interval. The sample was randomly divided into two groups to test parameter invariance, and item parameters were estimated through these groups. Findings support that item parameter invariance was attained.

In scale development or adaptation studies and studies in which measurement tools that intend to measure psychological characteristics are used (in particular for measurement tools used for measuring affective traits), when, in general, evaluating whether measurement tool provides factor structure, analysis of properties of item response categories is often ignored. However, rating level and psychometric properties of item response categories are also important for determining to what extent the measurement tool represents the construct it intends to measure. At this point, the fact that category threshold values are in acceptable intervals for each item and that observed category threshold values are comparable across items indicates that the information obtained from the items can be used in the same way. In computing total scores, it is relatively important that the extent of comparability of a response to an item, for example, a response of 4, with a response of 4 given to another item or the extent of equivalence of the distance between responses of 3 and 4 in an item to the corresponding distance in another item. This study focused on these questions and highlighted the importance of computation of item parameters for measurement tools comprising items that use an ordinal rating scale. It is suggested that model-data fit and item parameters should be studied in detail using models like GRM for ordinal rating scales such as 3-point or 5-point scales.

It is possible to determine at which levels the scale provides more information by obtaining more in-depth information on ability levels upon provision of detailed information on the measurement tool. For future studies, it may be an option to incorporate additional items that will provide more information, particularly on the ability levels for which the scale provided little information. Moreover, ensuring model-data fit for a measurement tool scaling based on IRT allows the estimation of invariant parameters of the scale even if it is applied to different groups. This will provide valid and reliable measurement results in comparisons for the results of the same measurement tool applied to different study groups.

REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. doi: 10.1007/BF02293814

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*(3), 253-270. doi: 10.3102/10769986025003253

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

Caycho-Rodríguez, T., Vilca, L. W., Carbajal-León, C., White, M., Vivanco-Vidal, A., Saroli-Araníbar, D., ..., Moreta-Herrera, R. (2021). Coronavirus anxiety scale: New psychometric evidence for the Spanish version based on CFA and IRT models in a Peruvian sample. *Death Studies.* doi: 10.1080/07481187.2020.1865480

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. doi: 10.18637/jss.v048.i06

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*(4), 523-562. doi: 10.1207/S15327906MBR3604_03

Cho, S., Drasgow, F., & Cao, M. (2015). An investigation of emotional intelligence measures using item response theory. *Psychological Assessment, 27*(4), 1241-1252. doi: 10.1037/pas0000132

de Ayala, R. J., Dodd, B. G., & Koch, W. R. (1990, April). *A comparison of the partial credit and graded response model in computerized adaptive testing*. Paper presented at the AERA Annual Meeting. Boston.

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

Demirtaşlı, N., Yalçın, S., & Ayan, C. (2016). The development of irt based attitude scale towards educational measurement course. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1*(7), 133-144. doi: 10.21031/epod.43804

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

Ferrando, P., Lorenzo, U., & Molina, G. (2001). An item response theory analysis of response stability in personality measurement. *Applied Psychological Measurement, 25*(1), 3-17. doi: 10.1177/01466216010251001

Flannery, W. P., Reise, S. P., & Widaman, K. F. (1995). An item response theory analysis of the general and academic scales of the self-description questionnaire II. *Journal of Research in Personality, 29*(2), 168-188. doi: 10.1006/jrpe.1995.1010

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

Gray-Little, B., Williams, V., & Hancock, T. (1997). An item response theory analysis of the Rosenberg self - esteem scale. *Personality and Social Psychology Bulletin, 23*(5), 443-451. doi: 10.1177/0146167297235001

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* New York: Springer Science and Business Media.

Hattie, J. (1992). *Self-concept.* Hillsdale, NJ: Erlbaum.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. doi: 10.1080/10705519909540118

Kahraman, N., Akbaş, D., & Sözer, E. (2019). Bilişsel-olmayan öğrenme durum ve süreçlerini ölçme ve değerlendirmede boylamsal yaklaşımlar: Duygu Cetveli Alan Uygulaması örneği. *Bolu Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 19*(1), 257-269. doi: 10.17240/aibuefd.2019.19.43815-459831

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

Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7*(1), 15-32. doi: 10.1177/014662168300700104

Köse, İ. A. (2015). Aşamalı tepki modeli ve klasik test kuramı altında elde edilen test ve madde parametrelerinin karşılaştırılması. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 15*(2), 184-197. Retrieved from https://dergipark.org.tr/tr/download/article-file/17439

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.424.2811&rep=rep1&type=pdf

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi: 10.1002/j.2333-8504.1994.tb01618.x

Min, S., & Aryadoust, V. (2021). A systematic review of item response theory in language assessment: implications for the dimensionality of language ability. *Studies in Educational Evaluation, 68*. doi: 10.1016/j.stueduc.2020.100963

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176. doi: 10.1177/014662169201600206

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

Roskam, E. E. (1985). Current issues in item response theory. In E. E. Roskam (Ed.), *Measurement and personality assessment* (pp. 3-19). Amsterdam: North Holland.

Rubio, V. J., Aguado, D., Hontangas, P. M., & Hernandez, J. M. (2007). Psychometric properties of an emotional adjustment measure: an application of the Graded Response Model. *European Journal of Psychological Assessment, 23*(1), 39-46. doi: 10.1027/1015-5759.23.1.39

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monography No. 17). Retrieved from https://www.psychometricsociety.org/sites/main/files/file-attachments/mn17.pdf?1576606975

Samejima, F. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika, 23*, 17-35. doi: 10.2333/bhmk.23.17

Silvia, P. J. (2021). The self-reflection and insight scale: applying item response theory to craft an efficient short form. *Current Psychology*. doi: 10.1007/s12144-020-01299-7

Steiger, J. H., & Lind, J. M. (1980, May). *Statistically based tests for the number of common factors*. Paper presented in Psychometric Society. Iowa City.

Wang, W. C., Wilson, M., & Shih, C. L. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement, 43*(4), 335-353. doi: 10.1111/j.1745-3984.2006.00020.x

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070. doi: 10.1037/0022-3514.54.6.1063

Wu, M., & Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal, 18*(2), 93-113. doi: 10.1007/BF03217438

Yaşar, M., & Aybek, E. C. (2019). Üniversite öğrencileri için bir yılmazlık ölçeğinin geliştirilmesi: Madde tepki kuramı temelinde geçerlilik ve güvenilirlik çalışması. *İlköğretim Online, 18(4),* 1687-1699. Retrieved from https://ilkogretim-online.org/fulltext/218-1597121020.pdf?1618815938

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*(2), 125-145. doi: 10.1177/014662168400800201

Aynı Tepki Kategorilerine Sahip Likert Maddelerin Psikometrik Özelliklerinin Çok Kategorili Madde Tepki Kuramı Modelleri ile İncelenmesi

Giriş

Ölçme araçlarında yer alan çok kategorili (polytomous) Likert tipi puanlanan maddelere verilen cevaplar, bireylerin durumlarını en iyi tanımlayan kategorileri seçtikleri varsayımıyla, öznel değerlendirmelerine dayanmaktadır. Yapılan bu öznel değerlendirmelere göre bireyler, karar verme süreçlerinde farklı kriterlere göre durumlarını değerlendirerek cevap vermektedir. Bireyler arası bu öznel karar verme farklılıklarını tanımlamak ölçekleme için oldukça önemlidir. Öznel değerlendirmelerdeki bu varyansı yakalayabilmek için ölçme araçlarının uygun yöntem ve teknikler ile incelenmesi önemlidir (Wang, Wilson, & Shih, 2006). Çünkü bireylerin ölçek maddelerine verdiği tepkilerden en doğru ve kullanışlı bilgiler ortaya çıkarmak ölçme ve değerlendirmenin en temel amaçlarındandır (Wu & Adams, 2006). Ölçme modellerindeki yeni gelişmeler ve yaklaşımlar ile ölçme uygulamalarındaki hataların azaltılması, doğruluğun ve etkililiğin arttırılması hedeflenmektedir (Baker, Rounds, & Zevon, 2000). Bu bağlamda kişilik özellikleri gibi psikolojik yapıların değerlendirilmesinde kullanılan farklı cevap formatlarına sahip ölçme araçlarının psikometrik özelliklerinin değerlendirilmesi için çok kategorili Madde Tepki Kuramı (MTK) modelleri geliştirilmiştir. MTK modellerine göre ölçeklendirilen test ve ölçekler ile her bir maddenin her bir kategorisine cevap vermek için gerekli olan yetenek düzeyinin ayrı ayrı kestirimi sağlanmaktadır. Bu da bireysel farklılıkların ölçümü bağlamında daha güvenilir ve geçerli sonuçların elde edilmesine neden olmaktadır. Bir ölçme aracında kullanılan cevap formatının ölçmeye çalıştığı psikolojik gerçekliğe ne derece uygun olduğu, ölçme aracından elde edilen ölçümlerin geçerliğini de etkilemektedir (Baker ve diğerleri, 2000). Dolayısıyla kullanılan veriye uygun bir modelin seçilmesi sonuçlardan elde edilecek çıkarımların anlamlılığı için önem taşımaktadır.

Psikolojik özellikleri ölçen ölçme araçları genelde çok kategorili cevap formatına sahip maddelerden oluşmaktadır. Bu maddelerin incelenmesinde kullanılan çok kategorili puanlanan MTK modelleri, cevaplayıcının yetenek düzeyi ile belli bir kategoride tepki verme olasılığı arasında doğrusal olmayan ilişkiler kuran modellerdir (Embretson & Reise, 2000). Çok kategorili modeller, madde ağırlıklandırmalarında farklı madde ayırt edicilik değerlerinin kullanılması, her bir yetenek düzeyinde ölçme hatası kestiriminin yapılması ve birey ve maddeler için parametre değişmezliğinin elde

edilmesini sağlamaktadır (Lord, 1980). Bir ölçeğin ölçmeye çalıştığı yapının kendini temsil eden tepki kategorilerine nasıl ayrıldığı, o ölçeğin güvenirliğini etkilemektedir (Linacre, 2002). Eşit aralıklı veya sıralama düzeyi gibi farklı özelliklere sahip tepki kategorileri MTK içinde yer alan çok kategorili modeller ile incelenebilmektedir. Farklı özelliklere sahip madde cevap (tepki) kategorileri, Aşamalı Tepki Modeli (ATM; Samejima, 1969), Kısmi Puanlama Modeli (KPM; Embretson & Reise, 2000) ve Dereceli Ölçekleme Modeli (DÖM; Andrich, 1978) gibi ölçme modelleri ile incelenmekte ve model-veri uyumları değerlendirilmektedir. Bir modelin veriye uygunluğunun değerlendirilmesinin yanında, modelin yanıtların altında yatan psikolojik gerçekliğe (yani, yanıtların formatı) ne kadar uygun olduğuna dair temel gözlemlerin de dâhil edilmesinin önemi vurgulanmaktadır (Samejima, 1996). Bu nedenle, incelenen ölçme aracının kategorilerine ait özelliklerin neler olduğu (her madde için kategorilerin benzer bir sıraya sahip olup olmadığı) ve ölçmeye çalıştığı psikolojik yapıya ne denli uygun olduğunun belirlenmesi, elde edilen ölçme sonuçlarının güvenirlik ve geçerliği açısından önemlidir. Bu çalışmanın amacı pozitif duygu durumlarının ölçülmesi için geliştirilen 14 maddelik bir Pozitif Duygu Durum (PDD) ölçeğinin içerdiği derecelendirilmiş (1'den 5'e kadar) maddelerin tepki kategorilerini ve bu kategorilerin maddeler arası ne derece benzerlik/farklılık gösterdiğini incelemektir. Bu amaçla, PDD ölçeğinde yer alan çok kategorili puanlanan maddelerin madde parametrelerinin kestiriminin farklı modeller ile elde edilmesi, bu modeller için hesaplanan model-veri uyumunun incelenmesi ve duygu durumu ölçeği boyunca farklı yetenek düzeylerinde elde edilen ölçümlerin ölçme kesinliğinin karşılaştırmalı olarak değerlendirilmesi amaçlanmıştır. PDD ölçeğinin çok kategorili cevap formatına sahip olması ve çok kategorili modellerle cevaplama süreçleri arasındaki teorik ilişki dikkate alındığında, ölçekte yer alan maddelere verilen tepkileri belirlemede kullanılan kategoriler arası eşik (threshold) parametrelerinin maddeler içi sıralı olup olmadığı ATM, KPM ve DÖM ile çalışılmış ve bu modellerin her birinin öngördüğü koşullar üzerinden, maddeler için varsayılan kategori eşik parametrelerinin ölçekteki tüm maddeler için değişmezliği varsayımının geçerliliği, uygulamada bu ölçek için toplanan veriler kullanılarak incelenmiştir.

Yöntem

Bu çalışma, PDD ölçeği'nin psikometrik özelliklerinin MTK modellerine göre incelendiği karşılaştırmalı betimsel bir çalışmadır. Uygulama verisinde 326 gönüllü üniversite öğrencisi yer almaktadır. Bu çalışmada kullanılan veriler Duygu Cetveli Alan Uygulaması (Kahraman, Akbaş, & Sözer, 2019) olarak adlandırılan daha geniş kapsamlı bir çalışmadan gelmektedir. Çalışma verilerinin elde edildiği PDD ölçeği, bireylerden her madde için kendilerine verilen cevap kategorilerinden (*hiç veya çok az* için 1'den *çok* için 5'e kadar) birini işaretlemelerini isteyen 14 maddeden oluşmaktadır. Ölçekte yer alan 14 pozitif duygu durumu şu şekildedir: Mutlu, huzurlu, memnun, iletişime açık, anlayışlı, motive, dayanıklı, güçlü, özgüvenli, azimli, başarılı, iyimser, cesur ve enerjik. Verilerin analizi R (R Core Team, 2016) programında "mirt" paketi (Chalmers, 2012) kullanılarak gerçekleştirilmiştir. PDD ölçeğinden elde edilen verilerin analizinde ATM, KPM ve DÖM kullanılarak madde parametre kestirimleri yapılmıştır. Verilerin analiz aşamasında elde edilen betimleyici istatistikler (ortalama, standart sapma) ve madde-toplam test korelasyon değerleri (r_{ij}) incelenmiştir. Bununla birlikte ölçeğin faktör yapısı (tek boyutluluk varsayımı) Açımlayıcı Faktör Analizi (AFA), Doğrulayıcı Faktör Analizi (DFA) ve paralel analiz ile incelenmiştir. Ölçeğin güvenirlik katsayısı Cronbach's $\alpha = .92$ olarak belirlenmiştir.

Sonuç ve Tartışma

Madde Tepki Kuramı modellerinin temel varsayımları olan tek boyutluluk ve yerel bağımsızlık incelendiğinde, ölçeğin faktör yapısına ilişkin yapılan analizler sonucunda ölçeğin tek boyutlu bir yapıya sahip olduğu belirlenmiştir. Tek boyutluluk varsayımının sağlanması durumunda yerel bağımsızlık varsayımının da sağlanacağı çeşitli çalışmalar tarafından belirtilmiştir (Embretson ve Reise, 2000; Hambleton & Swaminathan, 1985). Bu noktada, çalışma kapsamında kullanılan maddelerin tek boyutlu bir yapı gösterdiği faktör analizi sonuçlarına göre doğrulanmıştır. Tek boyutluluğun sağlanması nedeniyle yerel bağımsızlık varsayımının da karşılandığı varsayılmıştır.

PDD ölçeğinde yer alan 14 madde, üç farklı çok kategorili MTK modeli kullanılarak analiz edilmiş, model-veri uyum istatistikleri ve her modele göre elde edilen madde bilgi miktarları incelenmiştir. Model-veri uyumu daha düşük AIC, BIC değerleri ve modellerden elde edilen -2loglikelihood değerlerinin çiftler halinde karşılaştırılması ile değerlendirilmiştir. Model-veri uyumu karşılaştırmalarına göre, ATM ve DÖM modellerinin veriye daha iyi uyum sağladığı gözlenmiştir. Madde ve toplam test bilgi miktarları ile farklı modellerden sağlanan marjinal güvenirlik değerleri incelendiğinde en fazla bilgi miktarının DÖM'den elde edildiği gözlenmiştir. Bununla birlikte en yüksek güvenirlik katsayısı .93 olarak ATM modelinden elde edilmiştir. Bu noktada -2loglikelihood, serbestlik dereceleri ve χ^2 değerlerine göre çiftler halinde model karşılaştırmaları yapılmıştır. İkili model karşılaştırmaları sonucunda ATM modelinin veriye daha iyi uyum sağladığı sonucuna ulaşılmıştır. ATM ile her madde için a_i parametresi (ayırt edicilik) ve 5'li tepki kategorileri için 4 eşik parametresi kestirilmiştir.

Aşamalı Tepki Modeli'ne göre elde edilen 14 maddeye ait a_i parametreleri 1.25 ve 2.66 değerleri arasında değişmektedir. Buna göre, maddelerin orta ve yüksek düzeyde ayırt edicilik değerlerine sahip olduğu görülmektedir. Analizde her madde için 4 kategori eşik parametresi kestirimi yapılmıştır. Yetenek ölçeği boyunca birinci kategori kesişim parametre değerleri -2 etrafında, ikinci kategori kesişim parametre değerleri -1, üçüncü kategori kesişim parametre değerleri 0 ve dördüncü kategori kesişim parametre değerleri 1.5 etrafında dağıldığı görülmektedir. Bu da ölçeğin, bireyleri yetenek ölçeği boyunca farklı yetenek düzeylerinde iyi bir şekilde ayırdığını göstermektedir. Test ve madde bilgi fonksiyonları ile ölçekten elde edilen bilginin ($-2 \leq \theta \leq +2$) aralığındaki yetenek düzeylerinde daha fazla olduğu görülmektedir. Parametre değişmezliğinin incelenmesi için örneklem tesadüfi olarak ikiye ayrılmış ve madde parametreleri bu gruplar üzerinden kestirilmiştir. Elde edilen bulgular, madde parametre değişmezliğinin sağlandığını desteklemektedir.

Ölçek geliştirme veya uyarlama çalışmalarında ve psikolojik özellikleri ölçmeye çalışan ölçme araçlarının kullanıldığı çalışmalarda (özellikle duyuşsal becerilerin ölçülmesinde kullanılan ölçme araçları için) genellikle ölçme aracının faktör yapısını sağlayıp sağlamadığı değerlendirilirken madde tepki kategorilerinin özelliklerinin incelenmesinin genelde ihmal edildiği görülmektedir. Oysaki ölçme aracının ölçmeye çalıştığı yapıyı ne derece temsil ettiğinin belirlenmesinde madde tepki kategorilerinin dereceleme düzeyi ve bu kategorilerin psikometrik özellikleri de önem taşımaktadır. Bu noktada, hesaplanan kategori eşik parametrelerinin her madde için kabul edilebilir aralıklarda yer alması ve gözlenen kategori eşik değerlerinin maddeler arası karşılaştırılabilir olması, maddelerden elde edilen bilginin aynı şekilde kullanılabilir olduğunu göstermektedir. Toplam puanların hesaplanmasında, bir maddeye verilen, örneğin, 4 cevabının, diğer bir maddeye verilen 4 cevabı ile ne kadar karşılaştırılabilir veya bir maddedeki 3 ile 4 cevabı arasındaki mesafenin bir diğer maddedeki aynı mesafeye ne kadar denk olduğu oldukça önemlidir. Mevcut çalışma bu sorulara odaklanmakta ve sıralama ölçeği kullanan maddelerden oluşan ölçme araçları için de madde parametrelerinin hesaplanmasının önemli olduğunun altını çizmektedir. Önerilen, 3'lü, 5'li gibi sıralı cevap kategorilerini kullanan maddelerden oluşan ölçekler için ATM gibi modeller ile model uyumu ve madde parametrelerinin detaylı bir biçimde çalışılmasıdır.

Ölçme aracına ilişkin ayrıntılı bilgilerin sağlanması ile yetenek düzeylerine ilişkin daha derinlemesine bilgiler elde edilerek ölçeğin hangi düzeylerde daha fazla bilgi sağladığı belirlenebilmektedir. Gelecek araştırmalarda kullanılacak ölçeğe, özellikle daha az bilgi sağladığı yetenek düzeyleri için daha fazla bilgi sağlayabilecek maddelerin eklenmesi düşünülebilir. Aynı zamanda, MTK'ya dayalı ölçekleme yapılan bir ölçme aracının model-veri uyumunun sağlanması ölçeğin farklı gruplarda uygulansa da değişmez parametre kestirimlerinin elde edilmesini sağlamaktadır. Bu durum, farklı çalışma gruplarına uygulanan aynı ölçme aracının sonuçlarına yönelik yapılacak karşılaştırmalarda geçerli ve güvenilir ölçme sonuçlarının elde edilmesini sağlayacaktır.