ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(3);321-343

The Number of Response Categories and the Reverse Scored Item Problem in Likert-Type Scales: A Study with the Rasch Model*

Likert Tipi Ölçeklerde Olumsuz Madde ve Kategori Sayısı Sorunu: Rasch Modeli ile Bir İnceleme

Mustafa İLHAN**   Neşe GÜLER***

Abstract

This study aims to address reverse scored item and the number of response categories problems in Likert-type scales. The Fear of Negative Evaluation Scale (FNES) and the Oxford Happiness Questionnaire (OHQ) were used as data collection tools. The data of the study were analyzed according to the Rasch model. It was found that the observed and expected test characteristic curves were largely overlapped, each of the three rating scales worked effectively, and the differences between response categories could be distinguished successfully by the participants in straightforward items. On the other hand, it was determined that there were significant differences between the observed and expected test characteristic curves in reverse scored items. According to the results the participants could not distinguish the response categories of the reverse scored items at three, five and seven-point rating versions of both scales. Hence, the reverse scored items were removed from the data file, and the analysis was repeated. The results revealed that item discrimination, reliability coefficients for person facet, separation ratios and Chi square values calculated for the facets of person and items were higher in five-pointed rating compared to three and seven pointed rating. Based on these results it can be said that the scale categories in reverse scored items could not be discriminated by responders at all type of rating, and that reverse scored items did not measure the same latent structure as straightforward items did.

Key Words: Likert type scale, reverse scored item, number of response categories, Rasch model

Öz

Bu araştırmada Likert tipi ölçeklerde olumsuz madde ve kategori sayısı sorununun ele alınması amaçlanmıştır. Çalışmada veri toplama aracı olarak Olumsuz Değerlendirilme Korkusu Ölçeği (ODKÖ) ile Oxford Mutluluk Ölçeği (OMÖ) kullanılmıştır. Araştırma kapsamında toplanan veriler Rasch modeline göre analiz edilmiştir. Analiz sonucunda; ODKÖ ile OMÖ'deki olumlu maddelerde gözlenen ve beklenen test karakteristik eğrilerinin büyük ölçüde örtüştüğü, her üç kategori sayısının da etkin bir biçimde çalıştığı ve ölçek kategorileri arasındaki farkların katılımcılar tarafından başarılı bir biçimde ayırt edildiği belirlenmiştir. Öte yandan olumsuz maddelerde gözlenen ile beklenen test karakteristik eğrileri arasında önemli farklılıklar olduğu saptanmıştır. Üç, beş ve yedili derecelendirmeden hangisi kullanılırsa kullanılsın, ODKÖ ile OMÖ'deki olumsuz maddelerde kategorilerin katılımcılar tarafından ayırt edilemediği tespit edilmiştir. Bu tespitin ardından olumsuz maddeler veri dosyasından çıkarılarak analiz tekrarlanmıştır. Elde edilen bulgular; madde ayırt ediciliği, birey yüzeyine ilişkin güvenirlik katsayısı ile birey ve madde yüzeyleri için hesaplanan ayırma oranı ve Ki Kare değerlerinin beşli derecelemede üçlü ve yedili derecelemeye göre daha yüksek olduğunu göstermiştir. Bu bulgular, üçlü, beşli ya da yedili derecelemeden hangisi kullanılırsa kullanılsın olumsuz maddelerde ölçek kategorilerinin cevaplayıcılar tarafından ayırt edilemediğine ve olumsuz maddelerin olumlu maddelerle aynı örtük yapıyı ölçmediğine işaret etmektedir.

Anahtar Kelimeler: Likert tipi ölçek, olumsuz madde, kategori sayısı, Rasch modeli

* It was orally presented at the 5th Congress on Measurement and Evaluation in Education and Psychology.

** Asst. Prof., Dicle University, Faculty of Education, Diyarbakir-Turkey. e-posta: mustafailhan21@gmail.com, ORCID ID: orcid.org/0000-0003-1804-002X

*** Assoc. Prof., Sakarya University, Faculty of Education, Sakarya-Turkey. e-posta: gnguler@gmail.com, ORCID ID: orcid.org/0000-0002-2836-3132

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 8, Sayı 3, Sonbahar 2017, 321-343.
Journal of Measurement and Evaluation in Education and Psychology, Vol. 8, Issue 3, Autumn 2017, 321-343.
Received: 12.06.2017
Accepted:14.09.2017
DOI: 10.21031/epod.321057

INTRODUCTION

Likert type scales were introduced to the literature by Rensis Likert in 1932 (Likert, 1932). A number of statements are presented to participants in such scales and they state the extent to which they agree with the statement on a continuum ranging between *strongly agree* and *strongly disagree* or between *very appropriate to me* and *not appropriate to me at all* (Erkuş, 2003). The development and implementation of Likert type scales are easier than other measurement tool (Ahlawat, 1985; Tezbaşaran, 1997; Tavşancıl, 2010). Therefore these scales are frequently used in research in social sciences, psychology and educational sciences (Adelson & McCoach, 2010; Chang, 1994). Due to their common use, a great number of studies were performed in relation to determining how changes in the format of Likert type scales affect the psychometric properties of measurements. How reverse scored items used in Likert type scales influence measurement results, and what the most appropriate number of response categories is in such scales are the two basic issues considered in the studies (Preston & Colman, 2000).

The Problem of Reverse Scored Item

One of the most fundamental problems in Likert type scales is about how useful the reverse scored items are in such scales, and about how validity and reliability are influenced by them. Reverse scored items are also known as negative items, and the high scores received from these items indicate that participants have the measured psychological structure at low levels (Chiorri, Anselmi & Robusto, 2009). Developing the scale in a manner as to include reverse scored as well as straightforward items is a commonly preferred practice in order to prevent response sets based on stereotyped judgement and to reduce bias in responses such as affirmation or agreement, and social desirability (Hooper, Arora, Martin & Mulis, 2013; Van Sonderen, Sanderman & Coyne, 2013). However, there can also be disadvantages in using straightforward and reverse scored items together in the scale (Zhang, Noor & Savalei, 2016). Hence, DeVellis (2003) calls attention to the fact that there can be a cost in including reverse scored items in the scale, and says that those items can cause confusion in responders. It is possible to come across with empirical studies overlapping with this view of DeVellis (2003) in the literature. For instance, Schrieheim and Hill (2003) conclude that reverse scored items which are used to raise validity by reducing acquiescence response bias cause decrease in validity on the contrary. Chamberlain and Cummings (1984) compared the reliability coefficients of the form containing straightforward and reverse scored items with those of the form containing only straightforward items. In consequence, they found that the reliability coefficient for the form containing only straightforward items was higher. Hooper et al (2013) found that reverse scored items made measurement models complicated and that they caused inclusion of variance irrelevant to the measured structure in results. Locker, Jokovic and Allison (2013) found that straightforward and reverse scored items were in different factors, and this result was considered as evidence for the fact that reverse scored items did not measure the same latent structure as straightforward items did. Conrad et al. (2004) analysed the restrictions of reverse scored items through Rasch analysis reaching similar conclusions, and they found that reverse scored items in the scale caused a decrease in model-data fit. In contrast to these studies pointing to the fact that reverse scored items did not work well and that they could threaten validity, Bergstrom and Luriz (1998) found that straightforward and reverse scored items measured the same structure, and that using these two types of items together was unobjectionable. Thus, it may be stated that there is no consistency between research findings on how reverse scored items influence validity and reliability.

Number of Response Categories Problem

The most frequently preferred number of response categories in Likert type scales is five-pointed rating recommended by Likert (1932) (Lozano, García-Cueto & Muñiz, 2008). However, a review of literature demonstrates that differing number of response categories can be used and that the issue of the most appropriate number is controversial. How the number of categories in the scale affects the averages and variance values of measurements, the distribution of the data, and the skewness and kurtosis coefficients (Dawes, 2008), and what categories participants preferred more (Preston & Colman, 2000) can be the subject matter of such controversy. In addition to that, the probable effects of the number of response categories on the validity and reliability of measurements are also the focus of those discussions (Turan, Şimşek & Aslan, 2015).

On examining the studies concerning how the number of response categories in a scale affects reliability, it was found that the majority of them (Aiken; 1983; Birkett, 1986; Chang, 1994; Cicchetti Shoinralter & Tyrer, 1985; Halpin, Halpin & Arbet, 1994; Jenkins & Taber, 1977; Lissitz & Green, 1975) were based on internal consistency reliability, and that relatively small portion of them (Boote, 1981; Oaster, 1989; Ramsay, 1973; Weng, 2004) were based on the effects of test-retest reliability. Those studies obtained inconsistent findings on the effects of response categories on both internal consistency reliability and test-retest reliability. Studies performed by Aiken (1983), Leung (2011), and Qasem, Almoshigah and Gupta (2014), for instance, found that the number of response categories in a scale did not have any effects on internal consistency reliability. Weng (2004), Lozano, Garcia-Cueto and Muniz (2008) and Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol and Coffman (2009), on the other hand, concluded that internal consistency reliability tended to rise in parallel to the increase in the number of response categories. While Matell and Jacoby (1971) found that the number of response categories did not affect test-retest reliability, Oaster (1989) and Weng (2004) reported that test-retest reliability rose with an increase in the number of response categories.

An examination of the studies analysing the correlations between the number of response categories in a scale and validity demonstrated that some of the studies tested the effects of the number of response categories on construct validity, and that some others sought answers to the question of whether or not criterion-based reliability differed according to the number of response categories used. Varied results were obtained in relation to the effects of the number of response categories on validity in the studies conducted. For example, Comrey and Montang (1982), and King, King and Klockars (1983) found that the rate of total variance explained and factor loads was higher and factor structure was more clear in seven-pointed rating than in two-pointing rating. Lozano, García-Cueto and Muñiz (2008) also found that the rate of variances explained in the factor analysis rose as the number of response categories in the scale increased; and the result was interpreted as that the increase in the number of response categories influenced validity in positive ways. In a similar vein, Tarkan (2015) concluded that factorial validity increased as the number of response categories in a scale increased. In contrast to these studies, Kim (1998) found in the study comparing three-pointed, five-pointed, seven-pointed and nine-pointed rating in terms of validity and reliability that validity was the lowest in three-pointed rating, medium in seven-pointed rating, and it was higher in five and nine-pointed rating. The study conducted by Maydeu-Olivares (2009) found that model-data fit decreased as the number of response categories in a scale increased.

There is no overlap between the findings for the effects on criterion-based validity as in the findings concerning the effects of the number of response categories on construct validity. In the study performed by Chang (1994) where four-pointed and six-pointed ratings were compared psychometrically, it was found that the number of response categories did not have any effects on criterion-based validity. In a similar way, Qasem, Almoshigah and Gupta (2014) also concluded that there were no significant differences between the criterion-based validity coefficients of the scales with two, three and five-pointing rating. Loken, Pirie, Virnig, Hinkle and Salmon (1987), on the other hand, point out that criterion-based validity is influenced by the number of response categories and that the criterion-based validity coefficients obtained from 11-pointed rating are higher than those calculated from three or four-pointed rating. However, Preston and Colman (2000) found that the criterion-based validity coefficients of the scales using two, three and four-pointed rating were lower, and the criterion-based validity coefficients of the scales with five or more categories were higher. Besides, it was also found that there were no statistically significant differences between the criterion-based validity coefficients of the scales having differing numbers of response categories.

The Purpose and Significance of the Study

This study has basically two purposes. First, it aims to determine the extent to which reverse scored items in Likert type scales are functional. In accordance with this aim, *i)* whether or not scale categories functioned in the same way in straightforward and reverse scored items was evaluated; *ii)* efforts were made to determine whether or not those items measured the same latent structure by comparing the test characteristic curves for those straightforward and reverse scored items. The operations mentioned were performed separately for Likert type scales having three, five and seven-pointed rating; and thus it was checked whether or not the functioning of the reverse scored items was influenced by the number of response categories in scales. Secondly, the study aims to exhibit the effects of the number of response categories used in Likert type scales on the psychometric properties of measurements. In line with this purpose, Likert type scales having three, five and seven-pointed rating were compared in terms of reliability and model-data fit. In this way, validity was not ignored while the effects of the number of response categories on reliability were being analysed. This is quite important to make study results more meaningful because, as it is also pointed out by Cronbach (1950), it is worthless to increase reliability on its own and validity should also be taken into consideration in order to be able state that a certain number of response categories raising reliability is appropriate.

This study differs from the previous studies in the literature in several aspects. Therefore, it is predicted that the study will contribute significantly to the relevant literature. Firstly, the number of response categories and reverse scored items are considered separately in the studies available in the literature. This situation leaves the question of whether the differences in the number of response categories in Likert type scales have the same effects on straightforward and reverse scored items unanswered. In other words, while the most appropriate number of response categories for Likert type scales was investigated in the studies in the literature, general evaluations were made based on the items, but the effects of the number of response categories on straightforward and reverse scored items were not tested separately. Because this current study considers the number of response categories and the problem of reverse scored items together, it will be possible here to reveal how the differences in the number of response categories influence measurement results separately for straightforward and reverse scored items. Hence, this study differs from other studies in the literature in this respect.

Secondly, all of the studies in the literature aiming to determine the most appropriate number of response categories in Likert type scales and the way reverse scored items used in those scales affect the psychometric properties of measurements were conducted in other cultures. No studies are available in relation to Turkish culture. Yet, cultural properties have significant effects on the responses given to Likert type items. The significant effects were shown in many studies in the literature. For instance, Bachman and O’Malley (1984) found that there were differences between sets of responses given by white and black individuals in the USA, and that the likelihood of the blacks to use the extreme points in Likert type scales was higher than that of whites. Stening and Everett (1984), however, found that the likelihood of Japanese people to use the middle points was higher than that of American or British people in answering the Likert type scales. Huri and Triandis (1989) compared the sets of responses given by Spanish and non-Spanish participants through a Likert type scale of five and ten-point rating. Accordingly, it was shown that Spanish participants used the extreme points of the scales more frequently than the others. In addition to that, it was also found that the sets of extreme responses given by Spanish participants decreased on using 10-pointed rating, and that the answers given by non-Spanish participants were not influenced by the number of response categories in the scale. In a study conducted by Hofsteder (1998), it was found that the participants living in masculine cultures used extreme points more in answering the scale items, but that the participants coming from dominantly feminine cultures tended to use answers in the middle points. Lee, Jones, Mineyama and Zhang (2002) found that the likelihood of the Japanese and the Chinese to choose the middle points in Likert type scales was higher than that of Americans. Whether individualism or collectivism has priority in cultures is another issue influential in responses to Likert type scales. Johnson, Kulesa, Cho and Shavitt (2005) found that the probability of giving responses corresponding to the extreme points in the scale was higher in cultures where individualism is stressed. In cultures where collectivism is dominant, on the other hand, it was more probable to give responses in the middle points of the scale. Hooper et al. (2013) studied how reverse scored items in mathematics self-efficacy scale which was applied in Trends in International Mathematics and Science Study (TIMSS) worked. Accordingly, it was found that the effects of reverse scored items on model-data fit differed from one country to another. On considering these studies indicating that cultural factors influence the responses to Likert type scales, it may be said that it is difficult to generalise the conclusions reached in studies conducted in different cultures into Turkish culture. In this sense, it is hoped that such a study to be performed would contribute to literature.

Thirdly, although there are numerous studies in the literature trying to determine whether or not the psychometric properties of Likert type scale differ according to the number of response categories in a scale, they dominantly use methods based on Classical Test Theory (CTT). For this reason, there can be some points in which the above mentioned studies are inadequate in answering the question of how the psychometric properties of measurements are influenced by the number of response categories used in a scale. One such point is that CTT-based reliability calculations in Likert type scales are restricted to item reliability. According to CTT, variability observed between individuals all stems from the personal differences of participants. Therefore, a reliability coefficient is not reported for individuals in CTT (Taşdelen, Güler & Kaya Uyanık, 2015). In Rasch model, however, individuals in addition to items are considered as the sources of error. Thus, reliability coefficients are calculated for both items and persons in the Rasch analysis. In this context, using the Rasch model in this study will enable us to determine the effects of the number of response categories on reliability for both items and persons (Güler, İlhan, Güneyli & Demir). Another point in which CTT is inadequate in determining whether or not the adopted number of response categories is appropriate is that it does not inform one of category statistics. To put it in more clear terms, it is impossible in CTT to make an evaluation on whether or not participants can distinguish between the sequential points of scale categories. In Rasch analysis, on the other hand, a table of category statistics is reported, and it is possible to make inferences about how well participants can distinguish between scale categories by analysing the table.

Finally, it is evident that the studies in the literature investigating the problem of reverse scored items in Likert type scales use one single tool of measurement. This current study, however, employs two different scales. In the first scale to be used in this study (Happiness Scale), high scores indicate positive properties for participants whereas in the other scale (The Fear of Negative Evaluation Scale) high scores represent negative properties for participants. Thus, it will be possible in this study to determine whether or not reverse scored items function in the same way in scales where high scores represent desired properties (such as self-respect, self-efficacy and job satisfaction) and in measurement tools where high scores represent undesired properties (such as anxiety, burnout and stress). It is thought that the study is also original in this respect and that it will contribute to the literature.

METHOD

Study Group

The study was conducted with two different groups which were composed of 312 university students in total. The first group consisted of 197 participants, 112 (56.90%) of whom were female and 85 (43.10%) were male. The participants' ages ranged between 17 and 34 in this group, with average age of 21.66. The second group consisted of 115 participants, 64 (55.70 %) of whom were female and 51 (44.30 %) of whom were male. The ages in this group ranged between 18 and 32, with average age of 21.72.

Data Collection Tools

The Fear of Negative Evaluation Scale (FNES) and The Oxford Happiness Questionnaire-Short Form (OHQ-S) were used as the tools of data collection.

The Fear of Negative Evaluation Scale (FNES)

The FNES, which was developed by Leary (1983), was adapted into Turkish by Çetin, Doğan and Sapmaz (2010). The original form of the scale, which was in five-pointed Likert type, contained 12 items. The Turkish version of the scale, however, it was found that the discrimination index of item four in the original form was negative. Therefore, the item was removed from the scale. Through explanatory factor analysis (EFA) performed with the remaining 11 items, a one-factor structure explaining 40.19% of the total variance was obtained. In consequence of confirmatory factor analysis (CFA), the fit indices for the one-factor model were found as: RMSEA=.062, NFI=.96, CFI=.98, IFI=.98, RFI=.95, GFI=.95 and AGFI=.92. It was determined that factor loads ranged between .44 and .78 in EFA, and between .37 and .74 in CFA. Internal consistency, split-half reliability, and test-retest reliability coefficients calculated for FNES were found to be .84, .83 and .82 respectively. Eight of the 11 items in the Turkish version of FNES were straightforward items containing statements for worries about fear of negative evaluation. The remaining three items were reverse scored items stating that there were no worries about fear of negative evaluation.

The Oxford Happiness Questionnaire-Short Form (OHQ-S)

OHQ-S was developed by Hills and Argyle (2002), and was adapted into Turkish by Doğan and Akıncı Çötok (2011). The scale is in six-pointed Likert type. It has eight items in its original form. However, item four in its Turkish version was found to have low discrimination index (.17). Thus, the item was removed from the scale, and validity and reliability studies were conducted with the remaining seven items. Following the EFA, a one-factor structure was obtained, as in the original form of the OHQ-S. It was found in this one-factor structure that the rate of explained variance was 39.74%, and that the factor loads of the items ranged between .53 and .72. The findings obtained in CFA showed that the one-factor structure of the Turkish version of the OHQ-S had adequate fit indices [χ^2/sd=2.77, RMSEA=.074, AGFI=.93, GFI=.97, NFI=.92, CFI=.95, IFI=.95 RMR=.044]. In consequence of reliability study, the internal consistency coefficient for the scale was found as .74, and test-retest reliability as .85. Five of the seven items in the Turkish form of OHQ-S were straightforward items indicating happiness whereas the remaining two were reverse scored items containing statements of unhappiness.

Data Collection

The data were collected in the spring semester of 2015-16 academic year. The data collection tools were administered to students in the classroom setting. Prior to the application, the participants were informed of the purpose of the research, and only volunteers were participated to the study. The FNES was applied to the first group, and the OHQ-S was applied to the second group in three, five and seven-pointed rating types at intervals of one week. Only extreme points were labelled in all three types of rating (*strongly disagree* → *strongly agree*), and the points between the two extremes were not labelled. The thought that a clear labelling cannot be made in such an approach as in three, five and seven-pointed rating (Şeker & Gençdoğan, 2006; Østerås et al., 2008) was influential in adopting such an approach. That is to say, differences can be observed in measurement results in Likert type scales depending on how clearly the points of a scale is labelled (Wyatt & Meyers, 1987). Therefore, it will not be possible to determine whether the findings are the results of differences in the number of response categories or of uncertainty of labelling in relation to the categories when labelling is used for all response categories in three, five and seven-pointed rating. Setting out from this fact, only extreme points were labelled in all three type of scales (in three, five and seven pointed rating).

Data Analysis

The obtained data were analysed through the Rasch model by using the FACETS package programme. Rasch model is a one-parameter model placed under the roof of item response theory (Baker, 2001). Each source of variability capable of influencing measurement results is called a facet in this model (Sudweeks, Reeveb & Bradshawc, 2004). Rasch model can be assessed under the titles of *two-facet* or *many-facet* according to the number of facets it contains. In the two-facet model, there are two sources of variability capable of influencing measurement results (Linacre, 2014) - namely, items and persons. In many-facet model, however, in addition to items and persons, there are also other sources of variability such as raters, or demographic properties for persons which can influence measurement results (Knoch & McNamara, 2015). Sources of variability capable of influencing measurement results are restricted to items and persons in this study. Therefore, two-facet Rasch model was used in the analysis of the data collected in this study. The Rasch analysis was carried out according to rating scale model and Joint Maximum Likelihood Estimation Method (Unconditional Maximum Likelihood Estimation-UCON). Rasch analysis outputs are composed of many tables and graphs such as category statistics, test characteristic curves, and measurement reports for the facets of item and person. The tables and graphs were analysed in accordance with the purpose of this study, and the analysis outputs on which each sub-purpose was based were shown in Table 1.

Table 1. Statistical Indicators Considered for each Sub-purpose of the Study

Sub-purposes	Statistical Indicators to be Considered	
To determine whether or not scale categories function in the same way in straightforward and reverse scored items	Table of Category Statistics	The statistical indicators were analysed for Likert type scales having three, five and seven-pointed rating separately to test the effects of the number of response categories on the functioning of reverse scored items.
To find whether or not straightforward and reverse scored items measure the same latent structure.	Test Characteristics Curve	
To find determine the effects of response categories on reliability	Reliability coefficients, separation ratio and Chi-square for the facets of item and person	
To demonstrate the effects of the number of response categories on validity.	Infit and outfit statistics showing the model-data fit	

RESULTS

In this part of the results of the study are presented. First, category statistics table was analysed so as to determine how actively three, five and seven-pointed rating worked. The category statistics for the straightforward items in FNES and OHQ-S are shown in Table 2, and the category statistics for reverse scored items are shown in Table 3.

Table 2. The Category Statistics for the Straightforward Items in FNES and OHQ-S

		Category	FNES				OHQ-S			
			Frequency and %	Avge Measure	Expected Measure	Outfit	Frequency and %	Avge Measure	Expected Measure	Outfit
Scaling	Three	1	848 (54%)	-1.44	-1.32	.9	81 (14%)	-.56	-.32	.8
		2	563 (36%)	-.70	-.82	.8	310 (54%)	.36	.40	.7
		3	165 (10%)	-.10	-.29	.7	184 (32%)	1.33	1.17	.9
	Five	1	465 (30%)	-.94	-.82	.8	60 (10%)	-.44	-.36	.8
		2	492 (31%)	-.65	-.64	.7	60 (10%)	-.20	-.06	.6
		3	237 (15%)	-.46	-.48	.8	146 (25%)	.14	.20	.6
		4	312 (20%)	-.16	-.31	.6	168 (29%)	.52	.48	.6
		5	70 (4%)	.07	-.12	.8	141 (25%)	.93	.82	.9
	Seven	1	637 (40%)	-.54	-.49	.9	64 (11%)	-.29	-.23	.9
		2	286 (18%)	-.49	-.42	.6	48 (8%)	-.18	-.11	.7
		3	204 (13%)	-.30	-.35	.5	57 (10%)	-.06	-.01	.8
		4	167 (11%)	-.23	-.29	.6	113 (20%)	.09	.09	.7
		5	113 (7%)	-.11	-.23	.5	101 (18%)	.23	.20	.6
		6	59 (4%)	-.07	-.17	.7	75 (13%)	.34	.32	.8
		7	110 (7%)	-.06	-.11	.8	117 (20%)	.50	.45	.9

Making at least 10 observations in each category of the scale (for instance in each of the categories 1, 2 and 3) is the first assumption to meet in order to be able to say that rating adopted in the scale works actively (Linacre, 2014). According to Table 2, there are at least 10 observations for the straightforward items in FNES and OHQ-S for each category of three, five and seven-pointed rating. The second assumption to meet is that average measurements increase monotonously (Linacre, 2014). According to the Table, the average measurements in all three, five and seven-pointed rating increase in parallel to the scale categories. In other words, there is a continuous increase in three-pointed rating as moving from category 1 to category 3, in five-pointed rating as moving from category 1 to category 5, and in seven-pointed rating as moving from category 1 to category 7. The fact that outfit statistics are within the interval of .5 and 1.5 indicates that rating on which the scale is based works well (Linacre, 2014). According to Table 2, the outfit statistics are in the .5 and 1.5 interval in all three types of rating. These findings mean that all assumptions are met to be able to say that the rating used in the scale works actively. Thus, it may be said that all three types of rating work properly with straightforward items in FNES and in OHQ-S. Having found this, the data in Table 3 were analysed so as to determine how the scale categories worked with reverse scored items.

Table 3. The Category Statistics for the Reverse Scored Items in FNES and OHQ-S

			FNES				OHQ-S			
		Category	Frequency and %	Avge Measure	Expected Measure	Outfit	Frequency and %	Avge Measure	Expected Measure	Outfit
Scaling	Three	1	155 (26%)	.12	-.35	1.8	106 (46%)	-1.14	-1.46	1.4
		2	264 (45%)	-.07*	.04	1.1	107 (47%)	-1.07	-.90	1.4
		3	172 (29%)	.17	.49	1.3	17 (7%)	-1.31*	-.42	1.9
	Five	1	47 (8%)	.23	-.14	1.8	89 (39%)	-.62	-.90	1.5
		2	189 (32%)	.24	.02	1.8	61 (27%)	-.76*	-.66	1.4
		3	104 (18%)	.28	.16	1.8	45 (20%)	-.74	-.46	2.4
		4	161 (27%)	.14*	.31	1.5	21 (9%)	-.24	-.26	1.1
		5	90 (15%)	-.03*	.47	2.0	14 (6%)	-.59*	-.06	2.6
	Seven	1	106 (18%)	.03	-.14	2.0	88 (38%)	-.45	-.60	1.3
		2	95 (16%)	-.08*	-.07	1.3	45 (20%)	-.51*	-.51	1.2
		3	84 (17%)	.07	-.01	1.5	34 (15%)	-.57*	-.42	2.2
		4	76 (13%)	.14	.05	1.0	26 (11%)	-.49	-.34	1.9
		5	41 (7%)	.18	.11	.6	20 (9%)	-.36	-.26	1.8
		6	50 (8%)	.06*	.16	1.5	6 (3%)	-.12	-.19	.7
		7	139 (24%)	.00*	.21	1.7	11 (5%)	-.43*	-.12	2.3

The symbol (*) in the table shows that the assumption that average measurements increase in parallel to the scale categories was violated.

According to Table 3, the assumption that there should be at least 10 observations in each scale category in three, five and seven-pointed rating is met. However, the assumptions that the average measurements increase along with scale categories and outfit statistics should be in the .5 and 1.5 interval are not met in any of the three, five and seven-pointed rating. Accordingly, it can be stated that the scale categories cannot be distinguished by participants in reverse scored items in FNES and OHQ-S, no matter which (three, five or seven-pointed) type of rating is used.

Following the category statistics, the test characteristics curves were analysed for FNES and OHQ-S in order to decide whether or not straightforward and reverse scored items measured the same latent structure. The test characteristic curves for straightforward items in FNES and OHQ-S are shown in Figure 1.

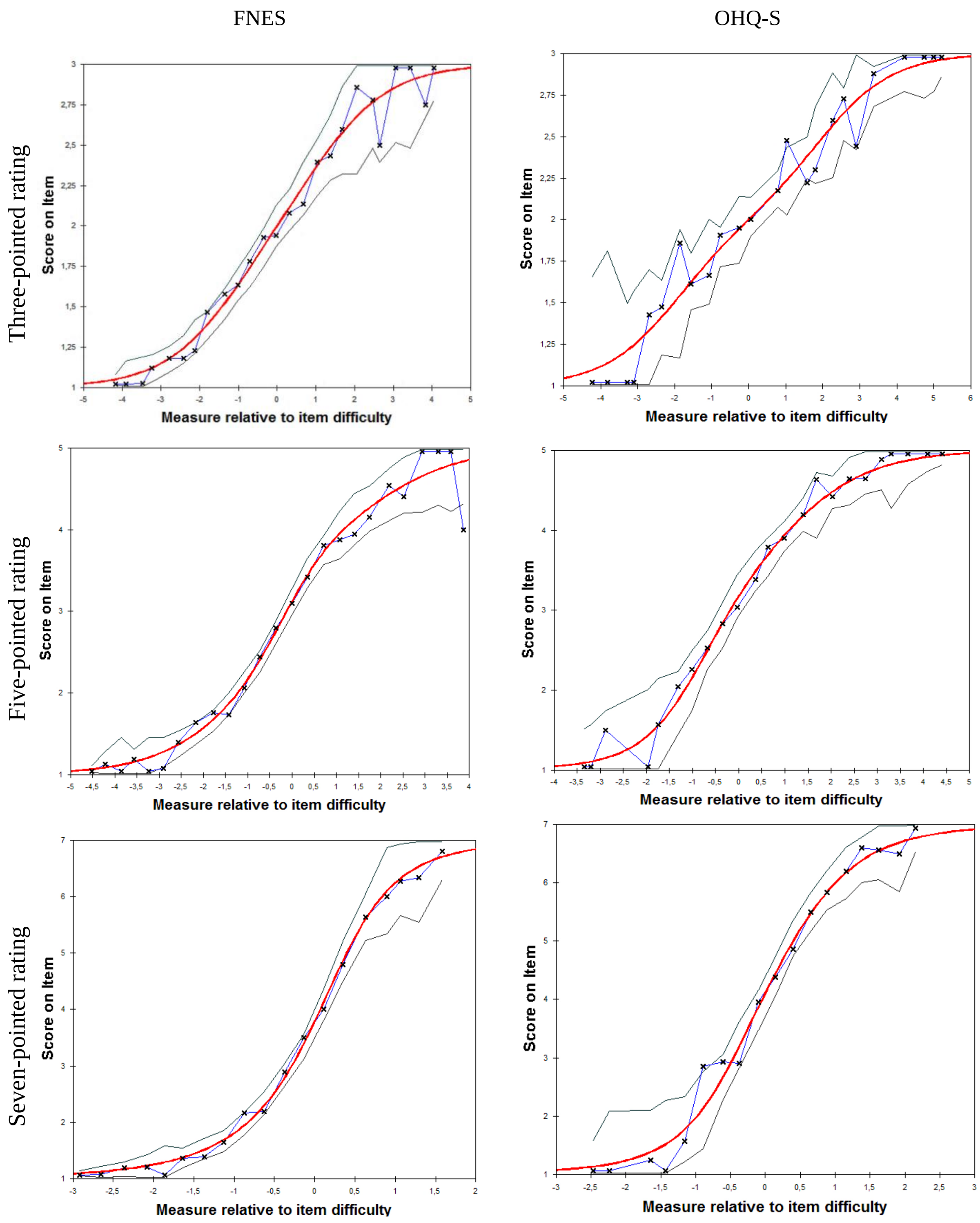

Figure 1. Test characteristic curves for straightforward items in FNES and OHQ-S.

As is clear from Figure 1, there are two lines – one of which is red and the other of which is blue on test characteristic curves. The red straight line represents the expected test characteristic curve while the blue line with crosses on it represents the observed test characteristic curve. The fact that there are no significant deviations between the expected and the observed test characteristic curves indicates model- data fit. Thus, it may be said that model-data fit is attained in all three types of rating for the straightforward items in FNES and OHQ-S. The fit shows that the straightforward items in FNES and OHQ-S can measure the latent structure which is targeted.

Test characteristic curves for the reverse scored items in FNES and OHQ-S are shown in Figure 2. According to Figure 2, there are important differences between the observed and the expected test characteristic curves for the reverse scored items in FNES and OHQ-S regardless of the type of rating. The differences show that the model-data fit is not attained in reverse scored items, and that therefore the reverse scored items in FNES and OHQ-S do not serve to measure the targeted structure. Accordingly, it may be said that the reverse scored items in FNES and OHQ-S do not measure the same latent structure as the straightforward items in FNES and OHQ-S do.

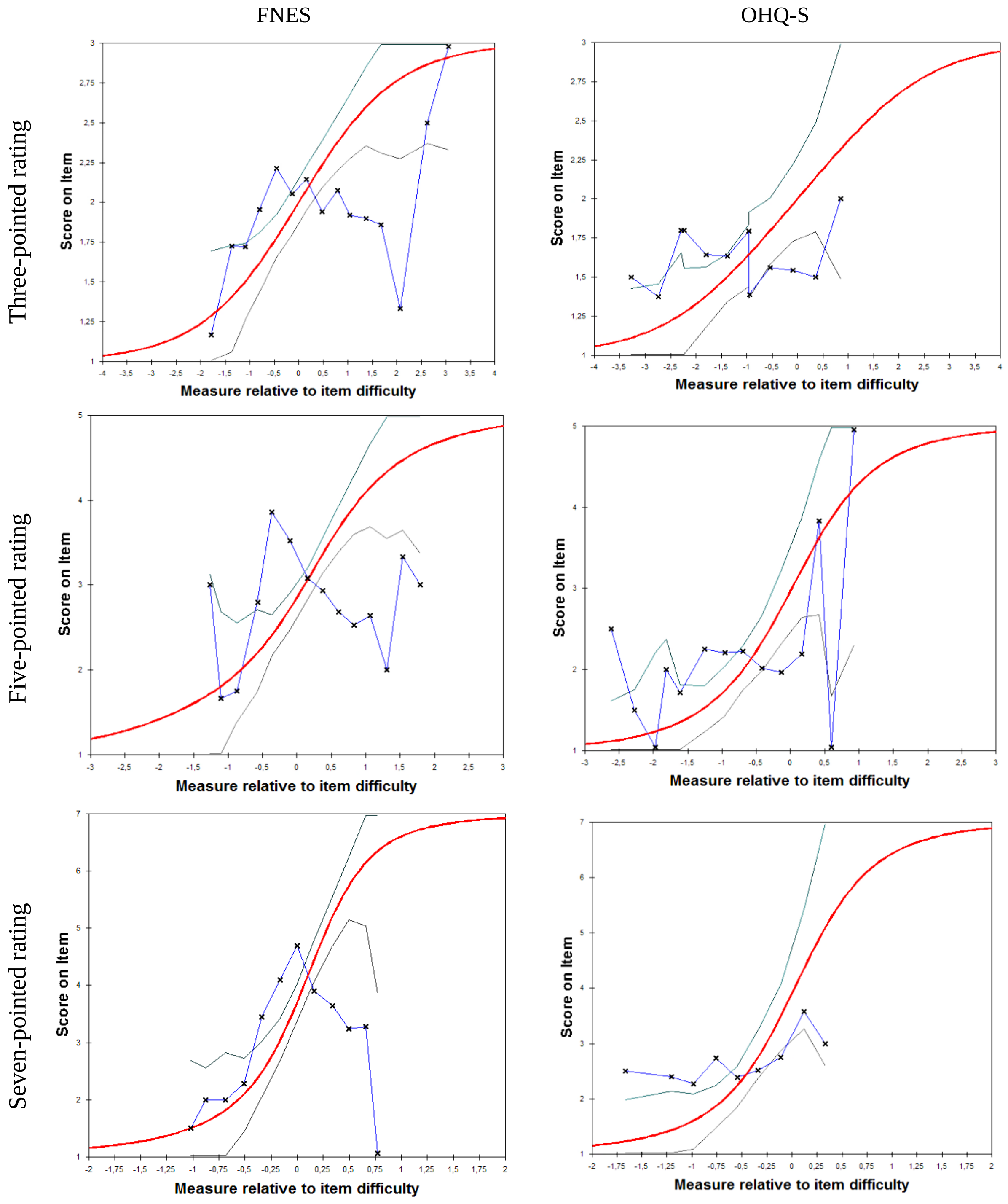

Figure 2. Test characteristic curves for reverse scored items in FNES and OHQ-S

Since category statistics reveal that the scale categories in reverse scored items cannot be distinguished by participants and since reverse scored items are found not to work in the same way as straightforward items according to test characteristic curves, the reverse scored items in FNES and OHQ-S were excluded while analysing the effects of rating type measurements on psychometric properties. Thus, reverse scored items were removed from the scale and Rasch analysis was repeated with straightforward items. Taking the confusing effects – which could stem from the fact that reverse scored items had psychometric properties different from straightforward items- under control was targeted in researching the effects of the number of response categories on validity and reliability with this study. Table 4 shows the measurement report for the facet of persons in FNES and OHQ-S in three, five and seven-pointed rating.

Table 4. The Measurement Report for the Facet of Person in FNES and OHQ-S in Three, Five and Seven-Pointed Rating

			FNES			OHQ-S		
			Measure	Infit	Outfit	Measure	Infit	Outfit
Scaling	Three	Mean	-1.58	1.02	.99	.80	.98	.94
		Standard Deviation	1.75	.58	.64	1.83	.79	.77
		Separation Ratio	1.57			1.53		
		Reliability	.71			.70		
		Chi-square (χ^2)	628.4			328.8		
		Degrees of Freedom	196			114		
	Five	Mean	-1.12	1.04	1.02	.63	1.00	.95
		Standard Deviation	1.68	.78	.78	1.38	.91	.82
		Separation Ratio	2.17			1.79		
		Reliability	.82			.76		
		Chi-square (χ^2)	897.5			359.6		
		Degrees of Freedom	196			114		
	Seven	Mean	-.94	1.05	1.02	.27	1.01	.98
		Standard Deviation	1.28	.91	.92	.85	.90	.84
		Separation Ratio	1.56			1.51		
		Reliability	.71			.70		
		Chi-square (χ^2)	743.9			285.5		
		Degrees of Freedom	196			114		

According to Table 4, there are no significant differences in the infit and outfit statistics calculated for the facet of person in FNES and OHQ-S. In all three types of rating, the infit and outfit statistics calculated for the facet of person are in the interval of .5 and 1.5- which is acceptable (Linacre, 2014). Accordingly, model-data fit can be said to be attained. Linacre (2014) states that the fit between a model and its data informs us of the validity of the data. Therefore, it may be stated that there are no significant differences between three, five and seven-pointed rating types in terms of the validity of data, and that the model-data fit is attained no matter what number of response categories is used.

An examination of separation ratios and reliability values in Table 4 shows that the values reported for three and seven-point rating types are very close. It was also found accordingly that the separation ratio for five-pointed rating and reliability values were higher than those calculated in three and seven-pointed rating. This separation ratio found for the facet of person in FNES and OHQ-S, reliability and Chi-square values show that the latent property to be measured is discriminated more successfully in five-pointed rating than in three or seven-pointed rating. After measurements for the facet of person, the measurement reports for the facet of items were analysed. The measurement report for the facet of item in FNES and OHQ-S in three, five and seven-pointed rating types are shown in Table 5.

Table 5. The Measurement Report for the Facet of Item in FNES and OHQ-S in Three, Five and Seven-pointed Rating Types

			FNES				OHQ-S			
			Measure	Infit	Outfit	Corr. PtBis	Measure	Infit	Outfit	Corr. PtBis
Scaling	Three	Mean	.00	1.00	.99	.53	.00	.99	.94	.44
		Standard Deviation	.62	.10	.15	.05	.98	.14	.15	.05
		Separation Ratio	4.13				4.83			
		Reliability	.94				.96			
		Chi-square (χ^2)	119.8				89.5			
		Degrees of Freedom	7				4			
	Five	Mean	.00	.99	1.02	.64	.00	.99	.95	.52
		Standard Deviation	.42	.22	.25	.06	.70	.24	.26	.07
		Separation Ratio	4.60				6.04			
		Reliability	.95				.97			
		Chi-square (χ^2)	149.8				119.0			
		Degrees of Freedom	7				4			
	Seven	Mean	.00	1.02	1.02	.62	.00	1.01	.98	.45
		Standard Deviation	.21	.13	.21	.04	.39	.10	.12	.06
		Separation Ratio	3.54				5.55			
		Reliability	.93				.97			
		Chi-square (χ^2)	96.0				107.1			
		Degrees of Freedom	7				4			

On examining Table 5, it is observed that the infit and outfit statistics calculated for the facet of item in FNES and OHQ-S are very close. In all three types of rating, the infit and outfit statistics are within the interval of .5 and 1.5 – which is recommended to be considered (Linacre, 2014). These values for fit statistics indicate that the model fits the data, and that the validity of the data is attained.

According to Table 5, the point biserial correlation values are available on the right of the columns of the infit and outfit statistics. These correlations are the counterpart for Pearson’s correlations (Linacre, 2014), and are considered as evidence for item discrimination (item validity). Point biserial coefficients are presented separately for each item and are also reported as an average coefficient for the overall scale in Rasch analysis outputs. However, the point biserial coefficients are not presented separately for each item in Table 5. They are shown as average values corresponding to the division of total correlation coefficients to the number of items in the scale. According to these average values, it was found that correlation coefficients for the five and seven-pointed rating in FNES were close. Almost no differences were found between point biserial correlation coefficients calculated for three and seven-pointed rating in OHQ-S. It was also found that the point biserial correlation coefficients calculated for five-pointed rating was higher than those calculated for three and seven-pointed rating. On considering the biserial correlation coefficients calculated in FNES and OHQ-S altogether, it is found that five-pointed rating yields higher correlation coefficients than three-pointed and seven-pointed rating. Therefore, it can be said that item discrimination rises when five-pointed rating is used instead of three or seven-pointed rating in Likert type scales.

On checking the reliability shown in Table 5, it is found that coefficients calculated for the three, five and seven-pointed types of rating in FNES and OHQ-S are almost the same. In other words, the number of response categories has no significant effects on item reliability. However, the separation ratio for the facet of item and the chi-square values differ according to the number of response categories. The highest values for the separation ratio and for the chi-square test in both FNES and PHQ-S were obtained in five-pointed rating. On comparing the separation ratio and chi-square test results for three-pointed and seven-pointed rating, it was found that the values for three-pointed rating were higher in FNES and that the values for seven-pointed rating were higher in OHQ-S.

Accordingly, although it looks impossible to make a clear inference as to in which type (three-pointed or seven-pointed) of rating items are discriminated better, it can be said that the items with differing levels of difficulty (items in which there are differences between the probabilities of participants' agreement and disagreement) are discriminated better in five-pointed rating.

DISCUSSION AND CONCLUSIONS

This study investigated reverse scored items and the number of response categories problem in Likert type scales. It was found accordingly that three, five and seven-pointed rating types all worked actively in straightforward items. Yet, it was also found that scale categories could not be distinguished by participants in reverse scored items no matter what number of categories was used. Not fulfilling the assumption of the *scale categories for reverse scored items were symmetrical and equi-distant* under the real conditions could be a cause for this problem (Locker, Jokovic & Allison, 2013). This finding of the research is supported by DeVellis' (2003) explanation that including reverse scored items in Likert type scales can present certain disadvantages. In the book entitled *Scale Development: Theory and Applications*, DeVellis (2003) states that participants can have confusion about what the responses *strongly agree* or *strongly disagree* mean while answering the reverse scored items in Likert Type scales. Such confusion can lead to not being able to distinguishing between scale categories in reverse scored items. Therefore, the findings of this research are aligned with the explanations made by DeVellis (2003). The findings obtained by Conrad et al (2004) are also similar to the ones obtained in this current study. In their study by Conrad et al (2004) by using Mississippi Scale for Posttraumatic Stress Disorder, they found that four out of six items violating model-data fit were reverse scored items and authors pointed out that the category statistics for those four items were problematic.

The results obtained by Bergstrom and Lunz (1998), however, differ from the ones obtained in this study. Bergstrom and Lunz (1998) administered the Job Satisfaction Scale of 36 items (19 straightforward and 17 reverse scored) to a study group containing 706 participants, and analysed the category statistics for the straightforward and reverse scored items in the scale. They found in consequence that the scale categories worked actively in both straightforward and reverse scored items. The inconsonance between results in Bergstrom and Lunz (1998) and in this study can be attributed to the different procedures followed in the two studies. The straightforward and reverse scored items in the data collection tools were analysed together in this study, and syntax was prepared on the basis of examples Linacre (2014) gave in the users' manual for FACETS programme. The numbers for the straightforward and reverse scored items in the scale, and a command to reverse score the scale categories in the reverse items were added to the syntax. In Bergstrom and Lunz (1998), on the other hand, straightforward and reverse scored items were included in two different data sets, and the analyses were performed separately for the two sets. That is to say, Bergstrom and Lunz (1998) did not analyse straightforward and reverse scored items as the items measuring the same structure, but they analysed the items as if they belonged to two different scales. This situation can be seen as the cause for the non-overlap between the results obtained in Begrstrom and Lunz (1998) and in this study.

This study found that the observed and the expected test characteristic curves in straightforward items overlapped to a large extent, but that there were significant deviations between the observed and the expected test characteristic curves in reverse scored items. Accordingly, it could be said that model-data fit was attained in straightforward items but that it was not attained in reverse scored items. This research finding showing that straightforward items served to measuring the intended latent structure but that reverse scored items failed to measure the same latent structure is supported by the findings obtained by Schriesheim and Eisenbach (1995), Meloni and Gana (2001), Conrad et al. (2004), Roszkowski and Soven (2010), Hooper et al. (2013), and Locker, Jokovic and Allison (2013). Schriesheim and Eisenbach (1995) found that the variance stemming from the property to be measured was higher in items containing positive statements than in items containing negative statements. In the study by Meloni and Gana (2001), the validity and reliability of the Italian version of Penn State Worry Questionnaire. In the study, high correlations were found between scores received from the overall scale and the straightforward items in the scale. On the other hand, the study found that the correlations between reverse scored items and the scores received from the overall scale were lower. Besides, it was also found in the above mentioned study that the correlations between scale items and the sores received from the scale of self-actualization - which was used for criterion validity in the study- were not significant in any of the reverse scored items. Meloni and Gana (2001) stated based on this finding that reverse scored items reduced the validity of a scale and that the psychometric properties of a scale would be improved by removing those items from the scale. Conrad et al (2004) analysed the limitations of reverse scored items through the Mississippi Scale for Posttraumatic Stress Disorder containing 35 items 25 of which were straightforward and 10 of which were reverse scored items. Accordingly, the study found that reverse scored items caused a reduction in model-data fit and that validity improved without loss of reliability when those items were removed from the scale. Roszkowski and Soven (2010) reported that reverse scored items had item total correlations lower than straightforward items, that those items constituted a separate factor in themselves and that Cronbach Alpha internal consistency coefficient rose on removing them. The same study also found that a distinctive increase occurred in item total correlations and a one-factor structure was obtained when reverse scored items were expressed as straightforward. Hooper et al (2013) found that reverse scored items caused reduction in model-data fit, they made a measurement model complex, and that they caused inclusion of variance irrelevant to the measured structure in measurement results. In a similar vein, Locker, Jokovic and Allison (2013) concluded that even though they were written to measure the same structure straightforward and reverse scored items were in differing factors, that items expressed in straightforward way had mutual correlations higher than those expressed in reverse scored way, and that there were significant differences between average scores received from straightforward items and the average scores received from reverse scored items. All of the studies mentioned above agree with the finding that including reverse scored items in Likert type scales would influence measurement results in negative ways. Demonstrating in many studies analysing the factor structures of Likert type scales composed of straightforward and reverse scored items (Benson & Hocevar 1985; Bolin & Dodder. 1990; Herche & Engelland 1996; Kelloway, Catano & Southwell 1992; Lai, 1994; McInerney, McInerney & Roche, 1994; Pilotte & Gable 1990; Rodebaugh et al., 2004; Spector, van Katwyk, Brannick & Chen, 1997) that reverse scored items constitute a separate factor in themselves also support the findings obtained in this research because –as Ibrahimoğlu (2001) states- gathering straightforward and reverse scored items under different factors in a scale could mean that reverse scored items cause the inclusion of variables other than the property to be measured in measurement results (cited in Weems, Onwuegbuzie & Lustig, 2003).

On examining the effects of the number of response categories used in a scale on item correlations, it was found that item discrimination was higher in five-pointed rating than in three and seven-pointed rating. This finding overlaps with the theoretical knowledge included in the article by Jacoby and Matell (1986) entitled "*Are scales with three-pointed rating good enough?"* Jacoby and Matell (1971) pointed out that a scale could not give detailed information if the number of response categories in a scale is too small and that the discrimination would decrease due to this. In their opinion, if the number of response categories is too big, on the other hand, reductions in item discrimination can occur due to the fact that participants cannot distinguish between different points of the scale. The fact that item discrimination calculated for five pointed rating in FNES and OHQ-S was higher than those calculated for three and seven-pointed rating is compatible with explanations made by Tezbaşaran (1997). In the book entitled *A Guide for Developing Likert Type Scales*, Tezbaşaran (1997) pointed out that three, five or seven-pointed rating could be used in Likert type scales, but that the most appropriate number of response categories was five.

Similar results were yielded in the three, five and seven-pointed forms of rating in FNES and OHQ-S in this study in terms of model-data fit. That is to say, it was found that the number of response categories had no significant effects on model-data fit. This finding is parallel to the one obtained by Daher, Ahmad, Winn and Selamat (2015). Daher et al (2015) analysed the data collected with three, four and six-pointed rating of Malay spiritual well-being scale according to the Rasch Model. In consequence, they found that the fit statistics calculated for all three rating types were similar and that the number of response categories did not have significant effects on model-data fit. Model-data fit is regarded as evidence for the validity of measurements in the Rasch analysis (Linacre, 2014). Therefore, the finding that the number of response categories had no considerable effects on model-data fit can be interpreted that valid measurements can be performed by using any of three, five and seven-pointed rating in a scale of items reflecting the structure to be measured.

It was found through Rasch analysis that the reported reliability for the facet of person separation ratio and Chi square rose on raising the number of response categories in the scale from three to five. This finding indicates that individuals at different levels of the latent structure to be measured are discriminated more effectively in five-pointed rating than in three-pointed rating. One of the basic factors determining how well individuals are discriminated in consequence of a measurement is the extent to which a scale is precise. As the number of response categories decreases, the sensitivity of a scale falls (Erkuş, 2012), and this fall in sensitivity can lead to a fall in reliability. Here, discrimination of individuals more effectively in five-pointed rating than in three-pointed rating can be explained with the fact that the sensitivity of measurements obtained from three-pointed rating is higher than that obtained from three-pointed rating. The study conducted by Ray (1980), which concludes that discrimination increases by raising the number of response categories from three to five, is also supportive of our findings.

The decrease in reliability values for the facet of person instead of increase when the number of response categories is raised from five to seven according to the findings reported in Rasch analysis can stem from participants' encountering problems in distinguishing between the categories in seven-pointed rating because the increase in the number of response categories in the scale can only increase sensitivity up to a certain point. And increasing the number of categories too much causes a fall in the perception of discrimination between categories (Erkuş, 2012), and as a result, this can influence reliability for the facet of person. At this point, the question of whether the number of response categories in seven-pointed rating is more than that human mind can distinguish between comes into mind. Büyüköztürk (2005) states that whether or not individuals can make discrimination carefully enough while responding to a scale of seven-pointed rating is a matter of discussion. Miller (1956), on the other hand, claims that human mind has the capacity to distinguish between seven different categories (Cited in Preston & Colman, 2000). The fact that the category statistics obtained in this study for seven-pointed rating met the assumptions necessary to say that the scale categories worked properly overlaps with Miller's (1956) claim. Accordingly, it can be said the number of response categories in seven-pointed rating is within the limits that human mind can distinguish between. In addition, since five-pointed rating is used more frequently than seven-pointed rating in Likert type scales (Lozano, García-Cueto & Muñiz, 2008), individuals can be more familiar with five-pointed rating and can discriminate between the differences in scale categories in five-pointed rating more effectively than in seven-pointed rating. This situation is thought to be the cause for higher reliability, separation ratio and Chi-square calculated for the facet of person in five-pointed rating than in seven-pointed rating.

It was found in this study that the reliability coefficients calculated for the facet of item in three, five and seven-pointed rating was almost equal. Reliability coefficients calculated for the facet of item in the Rasch analysis correspond to Cronbach Alpha internal consistency coefficients calculated in the CTT (Linacre, 2014). Therefore, it may be said that the studies demonstrating that the number of response categories have no significant effects on Cronbach Alpha internal consistency coefficients (Aiken, 1983; Leung, 2011; Matell & Jacoby, 1971; Preston & Colman, 2000; Qasem, Almoshigah & Gupta, 2014; Wong, Peng, Shi & Mao, 2011) are all supportive of our findings obtained in this study. In contrast to the above listed studies, there are also studies conflicting with those findings in the literature. The studies conducted by Weng (2004), Lozano, García-Cueto and Muñiz (2008), Maydeu-Olivares et al. (2009), Uyumaz (2013) and Tarka (2015) and reporting that Cronbach Alpha internal consistency coefficients rise as the number of response categories in a scale increases differ from this study in terms of their findings. According to Fabiola, Iwin, Jennifer and Zaira (2012), the inconsistencies observed in the research findings concerning the effects of the number of response categories on validity and reliability can stem from the differences in the measurement models (CTT or Item Response Theory) used. For this reason, it can be said that this study analysing the effects of the number of response categories on the psychometric properties of measurements through Rasch analysis is not adequate on its own to make clear inferences about the correlations between the number of response categories and item reliability. It is predicted that clearer statements will be made about how item reliability is affected by the number of response categories with an increase in the number of studies to be made through methods based on item response theory.

This study found that the separation ratios and Chi square calculated for the facet of items in five-pointed rating were higher than the values for three and seven-pointed rating. Based on this finding, it can be said that the items with differing levels of difficulty (the likelihood of participants' agreement or disagreement) can be discriminated better in five-pointed rating than in three or seven-pointed rating. It is believed that this result is related with the sensitivity of the rating used the scale and with how effectively scale categories are discriminated by participants- as in the separation ratio and Chi square for the facet of person. As the measurement reports and category statistics for the facet of person in FNES and OHQ-S indicate, five-pointed rating can yield more sensitive measurements than three-pointed rating, and it is composed of response categories through which participants are discriminated more easily than seven-pointed rating. This property might have made five-pointed rating more effective in discriminating between items of differing difficulty than three and five-pointed rating. Compared to three-pointed and seven-pointed rating, five-pointed rating has higher separation ratio and chi square values for the facet of item- which is a finding compatible with all finding except for item reliability obtained in this study. The fact that there were differences in separation ratios and chi square in favour of five-pointed rating despite the absence of differences in item reliability between three, five and seven-pointed rating could be attributed to the fact that reliability, separation ratio and chi square were the measurements reported in different metrics. While item reliability can take on values between 0 and 1, separation ratio can take on values ranging between 1 and ∞, and Chi square can be ranged from 0 to ∞ (Sudweeks, Reeveb & Bradshawc, 2004). Therefore, it can be more difficult for some difficulties to be manifested in reliability coefficients than in separation ratio and Chi square values.

To sum up the conclusions reached in this study, it was found that the scale categories in reverse scored items could not be discriminated by responders no matter which type of rating (three, five or seven-pointed) was used, and that reverse scored items did not measure the same latent structure as straightforward items did. Considering these results showing that reverse scored items made measurement models more complicated, preparing Likert type scales having only straightforward items can be evaluated as an application which can improve the psychometric properties of measurements. This study also found that the number of response categories did not have any effects on model-data fit. On the other hand, category statistics, item discrimination, reliability coefficients and Chi square calculated for the facets of person and items demonstrated that five-pointed rating was more functional than three or seven-pointed rating. This result leads to the recommendation that five-pointed rather than three or seven-pointed rating should be preferred. Yet, the restrictions of the study limit the generalizability of the findings and they also require that the recommendation should be interpreted in the framework of these restrictions. The restrictions of the study and the recommendations to be made for further research in accordance with the restrictions are as in the following.

LIMITATIONS AND RECOMMENDATIONS

The first restriction of this study has to do with the properties of the study group. The study was conducted with a group composed of university students. The best number of response categories for a scale can differ according to participants' age and level of education (Adelson & McCoach, 2010; Fabiola et al. 2012; Tekindal, 2009). Therefore, it may be recommended that such a study be conducted with participants of different age groups and educational levels. Also replication of a similar study on different samples from Turkey will lend the generalization of the research findings to Turkish culture. The second restriction of the study is the way the reverse scored items in the scales used in this study are revealed. Reverse scored items can be formed by using words with opposite meaning as well as using negative prefixes (or suffixes) (Sonderen, Sanderman and Coyne, 2013). According to the results Swain, Weathers and Niedrich (2008) obtained by analysing approximately 2000 scale items, reverse scored items are stated by using negative prefixes or suffixes by 81%. Therefore, this study preferred the scales having reverse scored items expressed by using negative prefixes or suffixes. Because this situation restricted the generalizability of the research findings, it could be recommended that a similar study be performed by using scales with reverse scored items which are stated in words with opposite meanings. And finally, the data for this study were collected through FNES and OHQ-S, and the number of reverse scored items in both scales is about one third of straightforward items. Using equal number of straightforward and reverse scored items or more reverse scored items in prospective studies might contribute to the generalizability of the findings.

REFERENCES

Adelson, J. L., & McCoach, D. B. (2010). Measuring the mathematical attitudes of elementary students: The effects of a 4-point or 5-point Likert-type scale. *Educational and Psychological Measurement*, 70(5), 796-807. http://dx.doi.org/10.1177/0013164410366694

Ahlawat, K. S. (1985). On the negative valence items in self-report measures. *The Journal of General Psychology*, *112*(1), 89-99. http://dx.doi.org/10.1080/00221309.1985.9710992

Aiken, L. R. (1983). Number of response categories and statistics on a teacher rating scale. *Educational and Psychological Measurement*, *43*(2), 397-401. http://dx.doi.org/10.1177/001316448304300209

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *The Public Opinion Quarterly*, *48*(2), 491-509. http://dx.doi.org/10.1086/268845

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

Barnette, J. J. (1999, April). *Likert response alternative direction: SA to SD or SD to SA: Does it make a difference?* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada. Retrieved from http://eric.ed.gov/?id=ED429125

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement*, *22*(3), 231–240. http://dx.doi.org/10.1111/j.1745-3984.1985.tb01061.x

Bergstrom, B. A., & Lunz, M. E. (1998, April). *Rating scale analysis: Gauging the impact of positively and negatively worded items*. Paper presented at the Annual Meeting of the American Educational Research Association. San Diego, CA. Retrieved from http://files.eric.ed.gov/fulltext/ED423289.pdf

Birkett, N. J. (1986). *Selecting the number of response categories for a Likert-type scale*. Retrieved from http://www.amstat.org/sections/srms/Proceedings/papers/1986_091.pdf

Bolin, B. L., & Dodder, R. A. (1990). The affect balance scale in an American college population. *The Journal of Social Psychology*, *130*(6), 839-40. http://dx.doi.org/10.1080/00224545.1990.9924639

Büyüköztürk, Ş. (2005). Anket geliştirme. *Türk Eğitim Bilimleri Dergisi, 3*(2), 133-151. Retrieved from http://www.tebd.gazi.edu.tr/index.php/tebd/article/view/315/297

Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of inter-rater reliability: A Monte-Carlo investigation. *Applied Psychological Measurement, 9*(1), 31-36. http://dx.doi.org/10.1177/014662168500900103

Chamberlain, V. M., & Cummings, M. N. (1984). Development of an instructor/course evaluation instrument. *College Student Journal, 18*(3), 246-250.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement*, *18*(3), 205-215. http://dx.doi.org/10.1177/014662169401800302

Chiorri, C., Anselmi, P., & Robusto, E. (2009). Reverse items are not opposites of straightforward items. In U. Savardi (Ed.), *The perception and cognition of contraries* (pp. 295-328). Milano: McGraw-Hill.

Comrey, A. L., & Montang, I. (1982). Comparison of factor analytic results with two choice and seven choice personality item formats. *Applied Psychological Measurement*, *6*(3), 285-289. http://dx.doi.org/10.1177/014662168200600304

Conrad, K. J., Wright, B. D., McKnight, P., McFall, M., Fontana A., & Rosenheck, R. (2004). Comparing traditional and Rasch analyses of the Mississippi PTSD scale: Revealing limitations of reverse-scored items. *Journal of Applied Measurement, 5*(1), 15-30. Retrieved from https://www.academia.edu/2832927/Comparing_traditional_and_Rasch_analyses_of_the_Mississippi_PTSD_scale_Revealing_limitations_of_reverse-scored_items

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*(1), 3-31. http://dx.doi.org/10.1177/001316445001000101

Çetin, B., Doğan, T. ve Sapmaz, F. (2010). Olumsuz değerlendirilme korkusu ölçeği kısa formu’nun Türkçe uyarlaması: Geçerlik ve güvenirlik çalışması. *Eğitim ve Bilim, 35*(156), 205-216.

Daher, A. M., Ahmad, S. H., Winn, T., & Selamat, M. I. (2015). Impact of rating scale categories on reliability and fit statistics of the Malay spiritual well-being scale using Rasch analysis. *Malaysian Journal of Medical Sciences, 22*(3), 48-55. Retrieved from http://www.bioline.org.br/pdf?mj15032

Dawes, J. (2007). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research, 50*(1), 61-77. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.417.9488&rep=rep1&type=pdf

DeVellis, R. F. (2003). *Scale development: Theory and applications*. Newbury Park: Sage.

Doğan, T. ve Akıncı Çötok, N. (2011). Oxford mutluluk ölçeği kısa formunun Türkçe uyarlaması: Geçerlik ve güvenirlik çalışması. *Türk Psikolojik Danışma ve Rehberlik Dergisi, 4*(36), 165-172. Retrieved from http://dergipark.ulakbim.gov.tr/tpdrd/article/view/1058000176/1058000178

Erkuş, A. (2003). *Psikometri üzerine yazılar*. Ankara: Türk Psikologlar Derneği Yazıları.

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme-I*. Ankara: Pegem Akademi.

Fabiola, G. B., Iwin, L., Jennifer, L. M., & Zaira, V. V. (2012). The effect of the number of answer choices on the psychometric properties of stress measurement in an ınstrument applied to children. *Evaluar, 12* 43-59. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar/article/download/4694/4488

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi- square fit indices in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 4*(2), 108-120, http://dx.doi.org/10.1080/10705519709540064

Güler, N., İlhan, M., Güneyli, A., & Demir, S. (2017). An evaluation of the psychometric properties of three different forms of Daly and Miller’s writing apprehension test through Rasch analysis. *Educational Sciences: Theory & Practice, 17*(3), 721-744. http://dx.doi.org/10.12738/estp.2017.3.0051

Halpin, G., Halpin, G., & Arbet, S. (1994). Effects of number and type of response choices on internal consistency reliability. *Perceptual and Motor Skills, 79*(2), 928-930. http://dx.doi.org/10.2466/pms.1994.79.2.928

Herche, J., & Engelland, B. (1996). Reversed-polarity items and scale unidimensionality. *Journal of the Academy of Marketing Science, 24*(4), 366-374. http://dx.doi.org/10.1177/0092070396244007

Hofstede, G. (1998). *Masculinity and femininity: The taboo dimension of national cultures*. Thousand Oaks, CA: Sage.

Hooper, M., Arora, A., Martin, M. O., & Mullis, I. V. S., (2013, June). *Examining the behavior of “reverse directional” items in the TIMSS 2011 context questionnaire scales*. Paper Presented at the 5th IEA International Research Conference. National Institute of Education, Nanyang Technological University, Singapore. Retrieved from http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2013/Papers/IRC-2013_Hooper_etal.pdf

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*(3), 296-309. http://dx.doi.org/10.1177/0022022189203004

Ibrahim, A. M. (2001). Differential responding to positive and negative items: The case of a negative item in a questionnaire for course and faculty evaluation. *Psychological Reports, 88*(2), 497-500. http://dx.doi.org/10.2466/pr0.2001.88.2.497

Jacoby, J., & Matell, M. S. (1971). Three-point likert scales are good enough. *Journal of Marketing Research, 8*, 495-500. Retrieved from https://www.jstor.org/stable/pdf/3150242.pdf?_=1472027712885

Jenkins, G. D., & Taber, T. D. (1977). A Monte-Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62*(4), 392-398. http://dx.doi.org/10.1037/0021-9010.62.4.392

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*(2), 264-277. http://dx.doi.org/10.1177/0022022104272905

Kelloway, E. K., Catano, V. M., & Southwell, R. R. (1992). The construct validity of union commitment: Development and dimensionality of a shorter scale. *Journal of Occupational and Organizational Psychology, 65*(3), 197-211. http://dx.doi.org/10.1111/j.2044-8325.1992.tb00498.x

Kim, K. H. (1998). An analysis of optimum number of response categories for Korean consumers. *Journal of Global Academy of Marketing Science, 1*(1), 61-86. http://dx.doi.org/10.1080/12297119.1998.9707386

King, L. A., King, D., & Klockars, A. J. (1983). Dichotomous and multipoint scales using bipolar adjectives. *Applied Psychological Measurement, 7*(2), 173-180. http://dx.doi.org/10.1177/014662168300700205

Knoch, U., & McNamara, T. (2015). Rasch analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 275–304). New York, NY: Routledge.

Lai, J. C. L. (1994). Differential predictive power of the positively versus the negatively worded items of the life orientation test. *Psychological Repors, 75*(3), 1507-1515. http://dx.doi.org/10.2466/pr0.1994.75.3f.1507

Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in Nursing & Health, 25*(4), 295-306. http://dx.doi.org/10.1002/nur.10041

Leung, S. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research, 37*(4), 412-421. http://dx.doi.org/10.1080/01488376.2011.580697

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 2-55.

Linacre, J. M. (2014). *A user's guide to FACETS Rasch-model computer programs.* Retrieved from http://www.winsteps.com/a/facets-manual.pdf

Lissitz, R. W., & Green, S. B. (1975). Effects of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60*(1), 10-13. http://dx.doi.org/10.1037/h0076268

Locker, D., Jokovic, A., & Allison, P. (2013). Direction of wording and responses to items in oral health-related quality of life questionnaires for children and their parents. *Community Dent Oral Epidemiol 35*(4), 255-262. http://dx.doi.org/10.1111/j.1600-0528.2007.00320.x

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 4(2), 73-79. http://dx.doi.org/10.1027/1614-2241.4.2.73

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement, 31*(3), 657-674. http://dx.doi.org/10.1177/001316447103100307

Maydeu-Olivares A., Kramp U., García-Forero C., Gallardo-Pujol, D., Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods, 41*(2), 295-308. http://dx.doi.org/10.3758/BRM.41.2.295

McInerney, V., McInerney, D., & Roche, L. (1994, July). Definitely not just another computer anxiety instrument: The development and validation of CALM: Computer anxiety and learning measure. Paper presented at the Annual Stress and Anxiety Research Conference, Madrid, Spain. Retrieved from http://files.eric.ed.gov/fulltext/ED386161.pdf

Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-type scales. *Perceptual and Motor Skills, 68*(2), 549-550. http://dx.doi.org/10.2466/pms.1989.68.2.549

Østerås, N., Gulbrandsen, P., Garratt, A., Benth, J. S., Dahl, F. A, Natvig, B., & Brage, S. (2008). A randomised comparison of a four- and a five-point scale version of the Norwegian function assessment scale. *Health and Quality of Life Outcomes, 6*(14), 1-9, http://dx.doi.org/10.1186/1477-7525-6-14

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement, 50*(3), 603-610. http://dx.doi.org/10.1177/0013164490503016

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1-15. http://dx.doi.org/10.1016/S0001-6918(99)00050-5

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika, 38*(4), 513-533. http://dx.doi.org/10.1007/BF02291492

Ray, J. (1980). How many answer categories should attitude and personality scales use? *South African Journal of Psychology, 10*, 53-54. Retrieved from http://jonjayray.tripod.com/howmany.html

Rodebaugh, T. L., Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L., & Rapee, R. M. (2004). More information from fewer questions: The factor structure and item properties of the original and brief Fear of Negative Evaluation Scale. *Psychological Assessment, 16*, 169-181. http://dx.doi.org/10.1037/1040-3590.16.2.169

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education, 35*(1), 113-130. http://dx.doi.org/10.1080/02602930802618344

Qasem, M., Almoshigah, T., & Gupta, S. (2014). The effect of number of alternatives on validity and reliability in Likert scale. *International journal of innovative research & studies, 3*(6), 324-333. http://dx.doi.org/10.13140/2.1.2237.2803

Schrieheim, C. A, & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement, 41*(4), 1101-1114. http://dx.doi.org/10.1177/001316448104100420

Spector, P. E, van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How Item characteristics can produce artifactual factors. *Journal of Management, 23*(5), 659-677. http://dx.doi.org/10.1016/S0149-2063(97)90020-9

Stening, B. W., & Everett, J. E. (1984). Response styles in a cross-cultural managerial study. *Journal of Social Psychology, 122*(2), 151-156. http://dx.doi.org/10.1080/00224545.1984.9713475

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239-261. http://dx.doi.org/10.1016/j.asw.2004.11.001

Swain S. D, Weathers D., Niedrich R. W. (2008) Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research 45*, 116-131. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=990097

Şeker, H. ve Gençdoğan, B. (2006). *Psikolojide ve eğitimde ölçme aracı geliştirme*. Ankara: Nobel.

Tarka, P. (2015). Likert scale and change in range of response categories vs. the factors extraction in EFA model. *Folia Oeconomica, 1*(311), 27-36. http://dx.doi.org/10.18778/0208- 6018.311.04

Taşdelen Teker, G., Güler, N., & Kaya Uyanık, G. (2015). Comparing the effectiveness of SPSS and EduG using different designs for generalizability theory. *Educational Sciences: Theory & Practice, 15*(3), 635-645. http://dx.doi.org/10.12738/estp.2015.3.2278

Tavşancıl, E. (2010). *Tutumların ölçülmesi ve SPSS ile veri analizi*. Ankara: Nobel.

Tekindal, S. (2009). *Duyuşsal özelliklerin ölçülmesi için araç oluşturma*. Ankara: Pegem Akademi.

Tezbaşaran, A. (1997). *Likert tipi ölçek hazırlama kılavuzu*. Ankara: Türk Psikologlar Derneği.

Turan, İ., Şimşek, Ü. ve Aslan, H. (2015). Eğitim araştırmalarında Likert ölçeği ve Likert tipi soruların kullanımı ve analizi. *Sakarya Üniversitesi Eğitim Fakültesi Dergisi, 30*, 186-203. Retrieved from http://dergipark.ulakbim.gov.tr/sakaefd/article/view/5000143504

Van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's Learn from cows in the rain. *PloS one, 8*(7), 1-7. http://dx.doi.org/10.1371/journal.pone.0068967

Weems, G. H., Onwuegbuzie, A. J., & Lustig, D. (2003). Profiles of respondents who respond inconsistently to positively- and negatively- worded items on rating scales. *Evaluation & Research in Education, 17*(1), 45-60. http://dx.doi.org/10.1080/14664200308668290

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956-972. http://dx.doi.org/10.1177/0013164404268674

Wong, C. S., Peng, K. Z., Shi J., & Mao, Y. (2011). Differences between odd number and even number response formats: Evidence from mainland Chinese respondents. *Asia Pacific Journal of Management, 28*(2), 379–399. http://dx.doi.org/10.1007/s10490-009-9143-6

Wyatt, R. C., & Meyers, L. S. (1987). Psychometric properties of four 5-point likert type response scales. *Educational and Psychological Measurement, 47*(1), 27-35. http://dx.doi.org/10.1177/0013164487471003

Zhang, X., Noor, R., & Savalei, V. (2016) Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PLoS ONE, 11*(6), 1-15. http://dx.doi.org/10.1371/journal.pone.0157795

UZUN ÖZET

Giriş

Bu çalışmanın iki temel amacı bulunmaktadır: Bunlardan ilki; Likert tipi ölçeklerdeki olumsuz maddelerin ne derece işlevsel olduğunun tespit edilmesidir. Bu amaç doğrultusunda araştırmada; *i)* olumlu ve olumsuz maddelerde ölçek kategorilerinin aynı şekilde çalışıp çalışmadığı incelenmiş, *ii)* olumlu ve olumsuz maddelere ait test karakteristik eğrileri karşılaştırılarak bu maddelerin aynı örtük yapıyı ölçüp ölçmedikleri belirlenmeye çalışılmıştır. Belirtilen işlemler üç, beş ve yedili derecelemeye sahip Likert tipi ölçekler için ayrı ayrı gerçekleştirilmiştir. Böylelikle olumsuz maddelerin işleyişinin ölçekteki kategori sayısından etkilenip etkilenmediği kontrol edilmiştir. Araştırmanın ikinci temel amacını Likert tipi ölçeklerde kullanılan kategori sayısının ölçümlerin psikometrik özellikleri üzerindeki etkisinin ortaya konulması oluşturmaktadır. Bu doğrultuda; üç, beş ve yedili derecelendirmeye sahip Likert tipi ölçekler güvenirlik ile model-veri uyumu açısından karşılaştırılmıştır. Bu sayede kategori sayısının güvenirlik üzerindeki etkisi incelenirken yapı geçerliği de göz ardı edilmemiştir. Bunun araştırma sonuçlarını anlamlı kılma adına oldukça önemli olduğu düşünülmektedir. Çünkü Cronbach'ın (1950) da belirttiği gibi yalnızca güvenirliği arttırmanın tek başına bir değeri bulunmamakta; güvenirliği arttıran bir kategori sayısının uygun olduğunun söylenebilmesi için geçerliğin de dikkate alınması gerekmektedir.

Yöntem

Araştırma, toplamda 312 üniversite öğrencisinden oluşan iki ayrı çalışma grubu üzerinde yürütülmüştür. Birinci çalışma grubunda 197 ve ikinci çalışma grubunda 115 katılımcı yer almıştır. Çalışmada veri toplama aracı olarak Olumsuz Değerlendirilme Korkusu Ölçeği (ODKÖ) ile Oxford Mutluluk Ölçeği-Kısa Formu (OMÖ-K) kullanılmıştır. Leary (1983) tarafından geliştirilip; Çetin, Doğan ve Sapmaz (2010) tarafından Türkçeye uyarlanan ODKÖ sekizi olumlu (olumsuz değerlendirilme korkusunu destekleyen) ve üçü olumsuz (olumsuz değerlendirilme korkusunu desteklemeyen) toplam 11 madde içermektedir. OMÖ-K ise Hills ve Argyle (2002) tarafından geliştirilmiş, Doğan ve Akıncı Çötok (2011) tarafından Türkçe'ye uyarlanmıştır. Bu ölçekte beşi olumlu ve ikisi olumsuz toplam yedi madde bulunmaktadır. Birinci çalışma grubundaki katılımcılara ODKÖ, ikinci çalışma grubundaki katılımcılara OMÖ-K üç, beş ve yedili dereceleme ile uygulanmıştır. Her üç derecelemede de kategorilerin yalnızca uç noktaları isimlendirilmiş (*Hiç Katılmıyorum* → *Tamamen Katılıyorum*); uç noktalar arasında kalan seçenekler için bir adlandırma kullanılmamıştır. Bu tür bir yaklaşımın benimsenmesinde, yedili derecelemede, üçlü ve beşli derecelemedeki kadar net bir isimlendirme yapılamayacağı düşüncesi etkili olmuştur. Şöyle ki, Likert tipi ölçeklerde ölçek noktalarının ne kadar net bir biçimde adlandırıldığına bağlı olarak ölçme sonuçlarında farklılıklar gözlenebilmektedir. Bu bakımdan üçlü, beşli ve yedili derecelemede tüm kategoriler için isimlendirme kullanılması halinde araştırma sonucunda ulaşılan bulguların gerçekten kategori sayısındaki farklılıktan mı; yoksa kategorilere ilişkin adlandırmaların aynı kesinlikte olmayışından mı kaynaklandığını belirlemek mümkün olmayacaktır. Bu noktadan hareketle çalışmada; her üç ölçek formunda da (hem üç, hem beş hem de yedili derecelemede) kategorilerin sadece uç noktaları isimlendirilmiştir. Araştırma kapsamında toplanan veriler FACETS paket programından yararlanılarak Rasch modeline göre analiz edilmiştir.

Sonuç ve Tartışma

Rasch analizinden elde edilen bulgular, ODKÖ ile OMÖ-K'deki düz puanlanan maddelerde gözlenen ve beklenen test karakteristik eğrilerinin büyük ölçüde örtüştüğünü, her üç dereceleme türünün de etkin bir biçimde çalıştığını ve ölçek kategorileri arasındaki farkların katılımcılar tarafından başarılı bir biçimde ayırt edildiğini ortaya koymuştur. Diğer taraftan ters puanlanan maddelerde gözlenen ile beklenen test karakteristik eğrileri arasında önemli farklılıklar olduğu ve ölçek kategorilerinin etkin bir biçimde çalışmadığı saptanmıştır. Üç, beş ve yedili derecelendirmeden hangisi kullanılırsa kullanılsın katılımcıların ters puanlanan maddelerde ölçek kategorilerini birbirinden ayırt edemediği belirlenmiştir. Olumsuz maddelerin ölçme modelini karmaşıklaştırdığı gösteren bu sonuçlar dikkate alındığında Likert tipi ölçeklerin sadece olumlu maddeleri içerecek şekilde hazırlanması, ölçümlerin psikometrik özelliklerinin iyileşmesine katkı sağlayacak bir uygulama olarak değerlendirilebilir. Araştırmada ayrıca, kategori sayısının model-veri uyumu üzerinde önemli bir etkisinin olmadığı tespit edilmiştir. Madde ayırt ediciliği, birey yüzeyine ilişkin güvenirlik katsayısı ile birey ve madde yüzeyleri için hesaplanan ayırma oranı ve Ki Kare değerlerinin ise beşli derecelemede üçlü ve yedili derecelemeye kıyasla daha yüksek olduğu sonucuna ulaşılmıştır. Bu sonuç, Likert tipi ölçeklerde üçlü veya yedili derecelemedense beşli derecelemenin tercih edilmesi şeklinde bir öneriyi beraberinde getirmektedir. Ancak, araştırmanın sınırlılıkları çalışmadan elde edilen bulguların genellenebilirliğini kısıtladığı gibi getirilen önerilerin de bu sınırlılıklar çerçevesinde yorumlanmasını gerekli kılmaktadır. Çalışmaya ilişkin sınırlılıklar ve bu sınırlılıklar doğrultusunda getirilebilecek ileri araştırma önerileri şu şekilde sıralanabilir.

Araştırmanın sınırlılıklarından ilki, çalışma grubunun özellikleri ile ilgilidir. Araştırma üniversite öğrencilerinden oluşan bir çalışma grubu üzerinde yürütülmüştür. Ölçek için en uygun kategori sayısı, katılımcıların yaşı ve eğitim düzeyine göre farklılık gösterebilmektedir. Dolayısıyla, bu tür bir çalışmanın farklı yaş gruplarından ve eğitim seviyelerinden katılımcılarla yapılması önerilebilir. Ayrıca, benzer bir çalışmanın Türkiye'den farklı örneklemler üzerinde tekrarlanması araştırma bulgularının Türk kültürüne genellenebilirliğini arttırması bakımından önem taşımaktadır. Çalışmada kullanılan ölçeklerdeki olumsuz maddelerin ifade ediliş şekilleri, araştırmaya ilişkin ikinci bir sınırlılıktır. Olumsuz maddeler, olumsuzluk ekleri (-me, -ma ve değil gibi) kullanılarak yazılabildiği gibi zıt anlamlı kelimeler kullanılarak da oluşturulabilmektedir. Bu çalışmada olumsuzluk ekleriyle ifade edilmiş olumsuz maddelerin yer aldığı ölçekler kullanılmıştır. Bu durum, olumsuz maddelerle ilgili araştırmada ulaşılan bulguların genellenebilirliğini kısıtladığından benzer bir çalışmanın zıt anlamlı kelimelerle oluşturulan olumsuz maddelerin yer aldığı ölçekler kullanılarak gerçekleştirilmesi önerilebilir. Son olarak bu araştırmanın verileri ODKÖ ve OMÖ ile toplanmış olup bu ölçeklerin her ikisinde de olumsuz madde sayısı olumlu madde sayısının yaklaşık üçte biri kadardır. Konu ile ilgili yapılacak ileri araştırmalarda olumlu ve olumsuz madde sayısının eşit ya da olumsuz maddelerin sayıca olumlu maddelerden fazla olduğu ölçeklerin kullanılması, çalışmadan ulaşılan bulguların genellenebilirliğine katkı sağlayabilir.