İ.Ü. Siyasal Bilgiler Fakültesi Dergisi
No: 55. (Ekim 2016). ss.1-20.

IMPACT OF THE NUMBER OF SCALE POINTS ON DATA CHARACTERISTICS AND RESPONDENTS' EVALUATIONS: AN EXPERIMENTAL DESIGN APPROACH USING 5- POINT' AND 7-POINT LIKERT-TYPE SCALES

Oylum KORKUT ALTUNA*
F. Müge ARSLAN**

Abstract
A remarkable deal of social research is based on data collected through the use of Likert-type scales. The optimal number of response categories in Likert-type scales has been subject to an academic debate for years. This article studies the differences between 5- and 7-point Likert-type scales using the SERVPERF Scale, which was developed by Cronin and Taylor in 1992, as the measuring instrument. A pretest-posttest control group experimental design was used to test whether the differently pointed response categories lead to any statistical differences in data characteristics, dimensional structure of the scale and data fit. Results do not show any statistically significant differences in terms of normality and reliability whereas different dimensional structures are achieved for the 5- and 7-point scale formats of SERVPERF using Exploratory Factor Analysis. ANCOVA results reveal that the number of response categories is not affective on the participants' evaluations of SERVPERF. The results of confirmatory factor analysis show that the best fit is achieved for the 7-point SERVPERF.

Keywords: Number Of Response Categories In Scales, Likert-Type Scale, Pretest-Posttest Control Group Experimental Design, ANCOVA, SERVPERF.

ÖLÇEK MADDE SAYISININ CEVAPLAYICILARIN DEĞERLENDİRMELERİ VE VERİ KARAKTERİSTİĞİ ÜZERİNDEKİ ETKİLERİ: 5'Lİ VE 7 'Lİ LİKERT TİPİ ÖLÇEKLER ARASINDAKİ FARKLILIKLARIN DENEYSEL TASARIM KULLANARAK İNCELENMESİ

Öz
Sosyal Bilimler alanında yürütülen araştırmaların birçoğu Likert tipi ölçek kullanılarak toplanan veriye dayanmaktadır. Likert tipi ölçeklerde kullanılacak cevap kategorisi sayısı yıllardır akademik tartışmalara konu olmaktadır. Bu çalışmada 5 maddeli ve 7

* Associate Professor, Department of Business Administration, Faculty of Political Sciences at Istanbul University, Istanbul, Turkey. oaltuna@istanbul.edu.tr

** Professor, Department of Business Administration at Marmara University, İstanbul, Turkey. mugearslan@marmara.edu.tr

Araştırma Makalesi - Gönderim Tarihi: 31.08.2016 - Kabul Tarihi: 22.09.2016.

maddeli Likert tipi ölçekler arasındaki farklılıklar Cronin ve Taylor tarafından 1972 yılında geliştirilen SERVPERF ölçeği kullanılarak incelenmektedir. İki farklı ölçek türü kullanılarak elde edilen verilerde veri karakteristiği, ölçeğin boyutsal yapısı ve uyum istatistikleri açısından fark olup olmadığı, ön- test / son – test kontrol gruplu deneysel tasarım kullanılarak test edilmektedir. Araştırma bulguları, normallik ve güvenilirlik açısından iki ölçek türü arasında fark olmadığını ortaya koyarken; elde edilen boyutsal yapılar arasında fark olduğu tespit edilmiştir. ANCOVA analizi sonucunda kullanılan ölçek türünün, cevaplayıcıların SERVPERF ölçeğine ilişkin değerlendirmeleri üzerinde istatistiksel olarak anlamlı bir etkiye sahip olmadığı bulgusuna ulaşılmıştır. Doğrulayıcı Faktör Analizi sonuçlarına göre, 7 maddeli SERVPERF ölçeğinin uyum istatistiği değerlerinin, 5 maddeli SERVPERF ölçeğinden daha anlamlı olduğu tespit edilmiştir.

Anahtar Kelimeler: Ölçeklerde Cevap Kategorisi Sayısı, Likert Tipi Ölçek, Ön-Test / Son – Test Kontrol Gruplu Deneysel Tasarım, ANCOVA, SERVPERF.

Introduction

In social research, rating scales are among the most widely used instruments that are used to measure respondents' perceptions, attitudes, opinions and/or evaluations. One of the major design-related issues that the researchers face during measurement is the number of response categories to be offered in the scale[1]. As known from literature, the number of response categories in a scale is one of the scale characteristics that are affective on the way people respond to such scales[2].

In current research studies, most of the rating scales that include the Likert-type scales comprise of either 5- or 7-point response categories[3]. Likewise, there are a number of textbooks on the subject that portray 5- or 7-point formats as the most common scale formats besides the 10- or 11-point scales which are also frequently used[4].

For years, there is an ongoing debate on the optimal number of response categories in a

[1] B.Weijters, E.Cabooter and N. Schillewaert, (2010). The Effect of Rating Scale Format on Response Styles: The Number of Response Categories and Response Category Labels. *International Journal of Research in Marketing*, 27, p.236.

[2] D.Weathers, S.Sharma and R.W.Niedrich, (2005), The Impact of the Number of Scale Points, Dispositional Factors and the Status Quo Decision Heuristic on Scale Reliability and Response Accuracy. *Journal of Business Research*, 58, p.1516.

[3] W.O.Bearden, R.G.Netmeyer and M.Mobley, (1993), *Handbook of Marketing Scales: Multi-item Measures for Marketing and Consumer Behavior Research*. Newbury Park, CA: Sage.; S.Choudhury and D.Bhattacharjee, (2014), Optimal Number of Scale Points in Likert Type Scales for Quantifying Compulsive Buying Behaviour. *Asian Journal of Management Research*, 4(3), 431-440.; C.C.Preston and A.M.Colman, (2000), Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power and Respondent Preferences. *Acta Psychologica*, 104, 1-15.

[4] J.Dawes, (2008), Do Data Characteristics Change According to the Number of Scale Points Used?, *International Journal of Market Research*, 50(1), p.62.

scale[5]. The aim of this study is to provide insights on the issue of determining the optimal number of response categories in a scale by investigating the differences between two sets of data using different category formats of SERVPERF (one set 5-point; the other set 7-point Likert-type scales) developed by Cronin and Taylor in 1992[6]. The two sets of data are assessed and compared in terms of data characteristics, dimensional structure of the scale and also fit of data. In this study, the impact of the number of response categories on participants' evaluations of SERVPERF is examined using a pretest-posttest control group experimental design where the number of scale points is manipulated between the two measures.

Literature Review

There is a wide range of research on the effects of variations in rating scale formats including differences in the number of response categories[7]. A noteworthy amount of research is based on the effects of the number of response categories on reliability[8] whereas another group of studies have investigated the effects of the number of response categories on validity along with reliability[9]. In addition, some studies have focused on the changes in the shape of data (skewness and kurtosis) when different scale formats are used[10].There are also a few studies that examine the relationship between the number of

[5] C.C.Preston and A.M.Colman, (2000), p.2.

[6] J.J.Jr. Cronin and A.S.Taylor, (1992), Measuring Service Quality: A Reexamination and an Extension. *Journal of Marketing*, 56(3), 243-253.

[7] C.C.Preston and A.M.Colman, 2000, p.6.

[8] A.W. Bendig, (1953). The Reliability of Self-ratings as a Function of the Amount of Verbal Anchoring and the Number of Categories on The Scale. *The Journal of Applied Psychology*, 37, 38-41.; A.W. Bendig, (1954), Reliability and The Number of Rating Scale Categories. *The Journal of Applied Psychology*, 38, 38-40.; G. Brown, R.E. Wilding and R.L. Coulter, (1991). Customer Evaluation of Retail Salespeople Using the SOCO Scale: A Replication Extension and Application. *Journal of the Academy of Marketing Science*, 9, 347-351.; D.V. Cicchetti, D. Showalter and P.J. Tyrer, (1985). The Effect of Number of Rating Scale Categories on Levels of Inter-rater Reliability: A Monte-Carlo Investigation. *Applied Psychological Measurement*, 9, 31-36.; E.P.Cox, (1980). The Optimal Number of Response Alternatives for a Scale: A Review. *Journal of Marketing Research,* 17, 407-422.; M.S. Matell and J. Jacoby, (1971). Is There an Optimal Number of Alternatives for Likert Scale Items? Study 1: Reliability and Validity. *Educational and Psychological Measurement*, 31, 657-674.; T.R.F. Oaster, (1989). Number of Alternatives per Choice Point and Stability of Likert-type Scales. *Perceptual and Motor Scales*, 68, 549-550.; J.O. Ramsay, (1973). The Effect of Number of Categories in Rating Scales on Precision of Estimation of Scale Values. *Psychometrika*, 38, 513-533.; P.M.Symonds, (1924), On the Loss of Reliability in Ratings Due to Coarseness of the Scale. *Journal of Experimental Psychology*, 7, 456-461.; L.J.Weng, (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

[9] L.Chang, (1994). A Psychometric Evaluation of Four-point and Six-point Likert-type Scales in Relation to Reliability and Validity. *Applied Psychological Measurement*, 18, 205-215.; M.S. Matell and J. Jacoby, (1971). B.Loken, P.Pirie, K.A.Virnig, R.L. Hinkle and C.T. Salmon, (1987), The Use of 0-10 Scales in Telephone Surveys. *Journal of the Market Research Society*, 29(3), 353-362.; C.C.Preston and A.M.Colman, (2000).

[10] J.Dawes, (2008); R.H.Finn, (1972) Effects of Some Variations in Rating Scale Characteristics on the Means and Reliabilities of Ratings. *Educational and Psychological Measurement*, 32(7), 255-265.

response categories and other factors such as respondent preferences[11], discriminating power and the amount of verbal anchoring[12].

Although there are studies suggesting that the maximum amount of information is obtained by using response categories of 20-points or over[13], researchers of other studies have stated that only marginal additional information is gained by increasing the number of response categories to more than seven[14]. In his comprehensive study Cox[15], mentions that there is no single optimal scale width that is appropriate for all circumstances but nevertheless the optimal scale width is generally between 5- to 9-points.

Regarding the optimal number of response categories in a scale, there are four major criteria that may be taken into account: discriminability, transmitted information, reliability and response accuracy[16]. According to these criteria, a scale with optimal number of categories provides the optimum discrimination and is capable of transmitting most of the information available from the respondents while showing high reliability scores and response accuracies[17].

Distinct from the research investigating the effects of response categories on the above-mentioned issues, Viswanathan, Sudman & Johnson[18] argue the use of the number of response categories that are meaningful to respondents rather than trying to maximize the discrimination power. Their findings show that although the number of response categories in a scale influences the responses to a scale by eliciting finer discriminations when the number of response categories increase, the number of meaningful categories for an attribute also influences attribute ratings[19].

In their study, Weathers et al. examine the effects of the respondent characteristics as well as the effects of the number of scale points on reliability and response accuracy and the mediating role of the status quo heuristic (SQH)[20]. Regarding the number of scale points,

[11] R.R. Jones, (1968). Differences in Response Consistency and Subjects' Preferences for Three Personality Inventory Response Formats. In Proceedings of the *76th Annual Convention of the American Psychological Association*, 247-248.; C.C.Preston and A.M.Colman, (2000).
[12] A.W. Bendig, (1953). B.Weijters, E.Cabooter and N. Schillewaert, (2010).
[13] W.R. Garner, (1960). Rating Scales, Discriminability and Information Transmission. *Psychological Review*, 67,343-352.
[14] P.M.Symonds, (1924). J.A.Green and V.R. Rao, 1970, Rating Scales and Information Recovery: How Many Scales and Response Categories to Use? *Journal of Marketing*, 34, 33-39.
[15] E.P. Cox, (1980) The Optimal Number of Response Alternatives for a Scale: A Review. *Journal of Marketing Research,* 17, 407-422.
[16] M.Lai, Y.Li, Yongjian and Y.Liu, (2010). Determining the Optimal Scale Width for a Rating Scale Using an Integrated Discrimination Fuction. *Measurement*, 43, p.1459.
[17] M.Lai, Y.Li, Yongjian and Y.Liu, (2010), p.1459-60.
[18] M.Viswanathan, S.Sudman, and M.Johnson, (2004). Maximum versus Meaningful Discrimination in Scale Response: Implications for Validity of Measurement of Consumer Perceptions About Products. *Journal of Business Research*, 57, 108-124.
[19] M.Viswanathan, S.Sudman, and M.Johnson, (2004), p.119.
[20] D.Weathers, S.Sharma and R.W.Niedrich, (2005).

their results show that as the number of response categories increases, the complexity of the choice task also increases leading to deterioration in response accuracy.

In another study, the authors investigate the effects of labeling used pertaining to response categories and the number of scale points on response styles[21]. According to their results, Net Acquiescence Response Style (NARS) was found to be higher whereas Extreme Response Style (ERS) and Misresponse to Reversed Items (MR) scores were lower in conditions where all response categories were labeled. Regarding the number of scale points, their results showed that the 7-point scales where labels are used only at the extremes (which are the most widely used scale types in marketing studies) increases the level of MR compared to 5-point scales where labels are used only at the extremes.

In this study the impact of the number of response categories on participants' evaluations of the scale is examined using a pretest-posttest control group experimental design where the number of scale points is manipulated (*5- and 7-point Likert-type scales*) between the two measures. In order to be able to make such a comparison a well-known and highly reliable and valid scale, SERVPERF, has been selected as the instrument in the experimental design.

The SERVPERF scale developed by Cronin and Taylor in 1992 is one of the most widely known and used scales used for measuring service quality. SERVPERF is a version of SERVQUAL in which the expectation component is discarded and only the performance component and comprises of 5 dimensions: tangibility, reliability, responsiveness, assurance and empathy[22]. As it is based on the "performance only" perspective, it operationalizes service quality as customers' evaluations of the service encounter[23]. There are empirical studies that evaluate the validity, reliability and methodological soundness of service quality scales and present evidence of the superiority of SERVPERF when compared to other quality measurement scales[24]. The performance of SERVPERF was validated in several studies, a number of which are in the fast food industry[25].

Based on previous literature discussed above concerning the optimal number of response categories in rating scales, it is expected that the change in the number of response

[21] B.Weijters, E.Cabooter and N. Schillewaert, (2010).

[22] J.Jr.Cronin and A.S.Taylor, (1992); S.K.Jain, S.K. and G.Gupta, (2004). Measuring Service Quality: SERVQUAL vs SERVPERF Scales. *The Journal for Decision Makers*, 29(2), p.28.

[23] F.A.Carillat, F.Jaramillo and J.P. Mulki, (2007). The Validity of the SERVQUAL and SERVPERF Scales: A Meta-analytical View of 17 Years of Research Across Five Continents. *International Journal of Service Industry Management*, 18(5), 472-490.

[24] E.Babakus and G.W.Boller, (1992), An Emprical Assessment of the SERVQUAL Scale. *Journal of Business Research*, 24(3), 253-268.; M.K.Brady, J.J.Jr. Cronin, and R.R. Brand, (2002).; S.K.Jain and G.Gupta, (2004); L.Zhou, (2004), A Dimension-specific Analysis of Performance-Only Measurement of Service Quality and Satisfaction in China's Retail Banking. *The Journal of Services Marketing*, 18(6/7), 534-546.

[25] H.Qin, V.R.Prybutok, V.R. and Q.Zhao. (2010). Perceived Service Quality in Fast-food Restaurants: Empirical Evidince from China. *International Journal of Quality & Reliability Management*, 27(4) , 424-437.

categories in SERVPERF will lead to a change in data characteristics, dimensional structure of the scale and data fit. Thus, the following hypotheses have been developed:

H1: Exposure to different number of scale response categories (5-point vs. 7-point) will result in statistically significant differences in data characteristics between the experimental and control groups' evaluations of SERVPERF.

H2: Exposure to different number of scale response categories (5-point vs. 7-point) will result in statistically significant differences between the experimental and control groups' evaluations of SERVPERF assuming that these groups do not show any statistically significant differences in their pretest evaluations.

H3: Exposure to different number of scale response categories (5-point vs. 7-point) will result in differences in the dimensional structure of SERVPERF between the experimental and control groups' evaluations of SERVPERF.

H4: An increase in the number of scale response categories (from 5-point to 7-point) will result in a better fit of data.

Methodology

In order to evaluate the differences regarding 5- and 7-point Likert-type scales, the research process of this study begins with a pre-test. As part of the main study, the data collected from the experimental and control groups in both measures were analyzed and compared in terms of means, standard deviations, alpha and test – retest reliability coefficients, dimensional structure and normality. Additionally, Analysis of Covariance (ANCOVA) was conducted in order to determine the impact of the number of response categories on participants' evaluations of SERVPERF. Eventually, Exploratory Factor Analysis (EFA) was used to examine the dimensional structure of SERVPERF for the two groups and Confirmatory Factor Analysis (CFA) was conducted in order to assess which version of SERVPERF (*5- or 7-point*) scale fit the data better.

Instrument.

SERVPERF was selected as the measurement instrument in this study. The 22 items of SERVPERF were translated into Turkish and back translated into English by a linguistic expert. Endpoint-only labeling at the extremes of the scale was applied in both 5- and 7- point Likert-type scales of SERVPERF used in this study (*1=Strongly Disagree, 5/7=Strongly Agree).*

Pre-Test.

A pre-test was undertaken with 45 undergraduate business students at the Faculty of Political Sciences, Istanbul University in order to determine the fast food restaurant to

be evaluated by the participants in the main study. For the pre-test a questionnaire with an open ended question (*Which are your favorite fast food restaurants that you most frequently visit - at least once in a three month period?*) was administered in class. The results revealed that the fast food restaurants that the students visited most frequently were Simit Sarayı, Burger King, Mc Donald's and KFC respectively. As a result of the pre-test, although Simit Sarayı was the most frequently visited fast food restaurant, Burger King was selected and used in the main study as it is a global brand providing standardized service in all of its restaurants.

Participants, Design and Procedures.

The study was conducted on 151 undergraduate business students at the Faculty of Political Sciences, Istanbul University. Among the total number of participants (N=151), the majority was found to be males (55%) who are freshmen (45.4%). 17.3% of the total sample were sophomores whereas 37.3% were juniors. As for Burger King Restaurant visiting behavior, 8% of the sample visited Burger King Restaurants at least once a week. 10.7% of the participants mentioned that they visited Burger King Restaurants twice a month whereas 41.3% had a visiting frequency of once a month. The rest of the sample (40%) visited Burger King Restaurants once in three months.

As a pretest-posttest control group experimental design was applied for the research, from the total number of students in the sample subjects were randomly assigned to two test units: the experimental group (N=78) and the control group (N=73). Both groups were administered a 5-point scale format of SERVPERF at the pre-treatment measurement. All participants were required to write their names and student id numbers on the surveys to be used later in the post-treatment measure. The number of response categories used in the scale was included in the study as the independent variable manipulated. In the post-treatment measurement, the experimental group was exposed to treatment and was administered the 7-point scale format of SERVPERF whereas the control group was administered the 5-point scale format of SERVPERF. The experimental design used in the study may be symbolized as:

EG: $R \quad O_1 \quad X \quad O_2$
CG: $R \quad O_3 \qquad O_4$

As internal validity is "the basic minimum that must be present in an experiment before any conclusion about treatment effects"[26], a number of precautions were taken during the design of the experiment in order to avoid extraneous variables that could violate the internal validity. First of all, the participants were randomly assigned to the two test units: the experimental and the control groups. None of the participants were informed about the aim of the study or the experimental design before the measurements. Although Simit Sarayı was rated as the most frequently visited fast food restaurant followed by Burger King in the pre-tests, Burger King was selected as the brand to be evaluated in terms of service quality criteria as it is a global brand and provides standardized service in all of

[26] N.K.Malhotra. (2010). *Marketing Research: An Applied Orientation*. Sixth Edition, Boston, MA: Pearson, p.255.

its restaurants. Only those students who visited Burger King Restaurants at least once in a three month period were included in the study. The surveys were administered only to the freshman, sophomore and junior students and not to senior students as at the time of the study, senior students were taking a lesson on marketing research and the scale types and experimental designs were among the topics discussed in class lectures. Senior students were not included in the study in order to avoid the maturation and selection bias effects. A time period of 4 weeks was allowed between the two measures (*pre and post-treatment*) as to set a precaution to history, maturation and interactive testing effects. A face to face survey method was applied in both measures in order to avoid instrumentation effects.

Findings and Discussion

Rescaling.

In order to be able to compare the two scale formats (5- and 7-point Likert-type) used in this study the data was rescaled. The rescaling method used by Dawes was adapted in this study[27]. According to the mentioned method, the 5-point scale end points of SERVPERF were anchored to the end points of the 7-point scale and also the mid-point of the 5-point scale was anchored to the 7-point scale. In other words, in order to rescale the 5-point version of SERVPERF to 7-points, 1 remained as 1; 5 was rescaled to 7 and the mid-point of 3 was anchored to 4 (the mid-point of the 7-point scale); the remaining scale values "were inserted at equal numerical intervals"[28]. Hence, the 5-point scale was rescaled as 1→1; 2→2,5; 3→4; 4→5,5; 5→7.

Data Characteristics According to Scale Formats.

Data characteristics regarding the means, standard deviations, alpha coefficients and test – retest reliability coefficients for both measures at the pre- and post-treatment for the two test units (experimental and control groups) are provided in Table 1.

Internal Structure.

Test-retest reliability was assessed by Pearson's correlation between scores of the same scale from two testing sessions[29]. The overall mean scores obtained from the 5- and 7-point scale formats of SERVPERF were used for analysis.

Table 1
Pre-treatment and Post-treatment Internal Structure Measures for Experimental & Control Groups

		Pre-Treatment			Post-Treatment			
Group	***n***	***Mean***	***SD***	***α***	***Mean***	***SD***	***α***	***r***
EG	78	4.30	0.73	0.848	4.05	0.83	0.875	0.60*
CG	73	4.32	0.66	0.826	4.32	0.68	0.850	0.55*

EG: Experimental Group **CG:** Control Group
*$p < 0.01$

[27] J.Dawes, (2008).
[28] J.Dawes, (2008), p.269.
[29] L.J.Weng, (2004), p.964.

As may be seen in Table 1, the lowest mean score is achieved for the 7-point SERVPERF used in the post-treatment measurement. Regarding the standard deviations, the highest score is calculated for the 7-point SERVPERF indicating greater individual variation on the scale among students[30]. These findings show some similarity to the findings of Weng as his research revealed an increase in the means and standard deviations when the number of response categories was increased. However, in this research the lowest mean was obtained for the post-treatment 7-point case.

Paired sample t-tests were used to compare the means between the correlated samples of the two test units. The results for the experimental group reveal that the mean scores from the two administrations of the SERVPERF show statistically significant differences [$t_{(76)} = 3.222$; $p<0.05$; $r=0.34$] showing a medium effect size[31] whereas the scale means do not show statistically significant differences for the control group between the pre- and post- treatment measures [$t_{(71)} = 0.013$; $p>0.05$]. Additionally, the results of the independent samples t-tests do not show any statistically significant differences between the experimental and the control groups regarding their evaluations of SERVPERF before the treatment (pre-treatment where 5-point scales were used for both groups). However, regarding the experimental and control groups, differences were found between the post-test results between the two groups [t□149□ = -2.177; $p<0.05$; $r=0.18$] showing a small effect size. Hence, no differences were found between the 5-point versions of SERVPERF, but statistical differences were found between the 5-point and 7-point versions of SERVPERF. Assessing these results which reveal that statistically significant differences exist only for the experimental group, it may be said that the number of response categories have a statistical significant effect on the evaluations of the participants of SERVPERF.

All items in both 5- and 7-point scale formats were found to be meeting the 0.40 criterion for item-total correlation[32]. As seen from Table 1, the reliability coefficients are all relatively high. In order to test for statistically significant differences between more than two related alpha coefficients Feldt's test (for samples greater than 99) and Fisher Bonett tests (for samples less than 100) are used[33]. In this study, although the sample sizes are 77 and 73 for the experimental group and control groups respectively, both of the tests (Feldt's test and Fisher Bonett test) were used. The results show that, none of the differences between the alpha coefficients were statistically significant [EG: (W=1.2459, p=.83), (z=0.9382; SE= .2343; p=.8259); CG: (W=1.1600, p=.73), (z=.6121; SE=.2425; p=.73)] meaning that no statistical differences exist for the reliability measurements for the two groups for pre-test and post-test comparisons. However, Table 1 shows that the 7-point SERVPERF

[30] L.J.Weng, (2004), p.964.

[31] J.Cohen. (1988). *Statistical Power Analysis for the Behavioural Sciences*. Second Edition, New York, NY: Academic Press.

[32] N.Osteras, P.Gulbrandsen, A.Garratt, J.S.Benth, F.A.Dahl, B.Natvig, B. and S.Brage, (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.

[33] C.C.Preston and A.M. Colman. (2000), p.6; D.J.Woodruff and L.S. (1986). Tests for Equality of Several Alpha Coefficients When Their Sample Estimates are Dependent. *Psychometrika*, 51, 393-413.

of the experimental post-test group has yielded the highest coefficient α value among the coefficient measurements.

As internal consistency reliability considers the degree of interrelatedness among individual items, in order to examine the stability of scale scores across occasions it is recommended that test-retest reliability is assessed[34]. Taking into consideration that the evaluation of internal consistency reliability alone is often referred to be inadequate, the effect of scale format on test-retest reliability in addition to internal consistency was assessed[35]. As may be seen in Table 1, the scores of each measurement were correlated and the coefficients of stability for both test units were above the threshold level of 0.50[36]. Fisher's *r*-to-*z* transformation test was used to see if the two correlations were significantly different from each other[37]. The difference between the test-retest reliability coefficients of the experimental and the control groups was found to be statistically non-significant (z=.45, p>.05). The higher test-retest reliability score is derived from the experimental group where the scale format was changed from a 5-point Likert-type to a 7-point Likert-type scale. Although no statistical difference exists taking into consideration the reliability scores, the findings suggest that the use of more response categories may be said to have an incremental increase on the reliability of SERVPERF. These findings seem to confirm the findings of Preston and Colman which reveal that although the reliability coefficients do not show statistically significant differences for the different response categories, the most reliable scores are derived from scales with 7-, 8-, 9-, or 10-response categories as compared to 2-, 3- and 4-point categories. The results are also in parallel with the findings of Symond and Cicchetti et al., which state that there is an increase in the reliability scores correlatively with the increase in the number of response categories in the scale[38].

Normality.

Normality was assessed using skewness and kurtosis analysis, Kolmogorov-Smirnov (KS) and Shapiro-Wilk tests**.** Variates having skewness and kurtosis values that are close to zero imply a closer approach to normality[39]. Although there are no clear cut guidelines for interpreting measures of skewness and kurtosis, in most research, data is considered to be approximately normal in shape if the skewness and kurtosis values are found to be between -1.0 to +1.0[40]. The skewness values regarding the overall mean scores for

[34] L.J. Weng. (2004), p.957.

[35] J.M.Cortina. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, 78, 98-104.; L.Crocker, L. and J.Algina, 1986, *Introduction to Classical and Modern Test Theory*. New York, NY: Holt, Rinehart & Winston. ; L.J. Weng (2004).

[36] L.Marlow, D.Inman and C. Shwery. (2005) To What Extent are Literacy Initiatives Being Supported: Important Questions for Administrators. *Reading Improvement*, 42(3), p.181.

[37] D.C.Howell. (1992). *Statistical Methods for Psychology*. Boston, MA: Duxbury Press, p.251; S.W. Huck. (2008). *Reading Statistics and Research*. Fifth Edition, Boston, MA: Pearson Education, Inc., p.216; C.C.Preston and A.M. Colman. (2000), p.6.

[38] P.M.Symonds. (1924); D.V. Cicchetti, D. Showalter and P.J. Tyrer, (1985).

[39] S.O. Leung, 2011, A Comparison of Psychometric Properties and Normality in 4-,5-,6 and 11-Point Likert Scales. *Journal of Social Service Research*, 37, p.417.

[40] S.W. Huck. (2008), p.29.

both the experimental and control groups regarding the pre- and post-treatment measures were examined. Results for the experimental group showed that both 5- and 7-point scale formats produced negative skewness scores *(Pre-treatment Measurement=-0.143; Post-treatment Measurement=-0.768)* within the -1.0 threshold indicating a relatively normal distribution. For the control group (*Pre-treatment Measurement=-0.331; Post-treatment Measurement=0.262)* the pre-treatment measurement produced negative skewness scores, whereas post-treatment measures produced positive skewness scores. Although the signs are different for the pre- and post- treatment cases, since the values are within the -1.0 to +1.0 range, this indicates that the data of the control group shows a relatively normal distribution.

Regarding kurtosis values, for the experiment group a negative kurtosis value was achieved for the pre-treatment case and a positive kurtosis value was achieved for the post-treatment case (EG: *Pre-treatment Measurement=-0.263; Post-treatment Measurement=0.907*). For the control group a positive kurtosis value was achieved for the pre-treatment case and a negative kurtosis value was achieved for the post-treatment case (CG: *Pre-treatment Measurement=0.116; Post-treatment Measurement=-0.147*). However, as the kurtosis values are within the -1.0 to +1.0 range, this indicates that the data of the control group shows a relatively normal distribution.

As may be seen from these results, although there are some minor differences between the skewness and kurtosis values concerning the pre-treatment and post-treatment measures for the experiment and control groups for both cases the data show normal distribution. For that reason, it could be said that the change in the number of response categories in SERVPERF do not cause a significant change in the normality of distribution. The results for the KS and SW tests that were used to conduct formal statistical assessment of normality are given in Table 2.

Table 2
Kolmogorov-Smirnov (KS) and Shapiro-Wilk (SW) Statistics, Degree of Freedom and Significant Values

Test Unit(measure)	*KS Statistics*[a]	*df*	*Sig*	*SW Statistics*	*df*	*Sig*
EG(pre-test)	.085	78	.200*	.987	78	.458*
EG(post-test)	.091	78	.200*	.962	78	.028
CG(pre-test)	.090	73	.200*	.982	73	.396*
CG(post-test)	.133	73	.007	.983	73	.505*

a Lilliefors Significance Correction
EG: Experimental Group **CG:** Control Group
*p > 0.05

As seen in Table 2, for both experiment and control groups, the results of the KS and SW tests are mostly insignificant ($p > 0.05$) indicating that the distributions for most of the measures are normal (except for the KS score of post treatment in the control group; and SW score of post treatment of the experimental group). Therefore it may be said that, the change in the number of response categories of SERVPERF leads to a minor change in the shape of the distribution, hence normal distribution. The results are similar to the findings of the study conducted by Doğan, Özkara, Yılmaz & Torlak in which the 5-, 7- and 11-point versions of three different scales were compared in terms of data characteristics where no differences were found in the shape of the distribution[41].

Taking into consideration the above given information, as there are almost no statistically significant differences for internal structure and normality tests between the 5- and 7-point scales, **H1 is rejected**.

The impact of the number of response categories on the participants' evaluations of SERVPERF.

Much of the rationale for the hypothesis of the covariance analysis has been developed in the previous sections. Based on previous literature discussed above concerning the optimal number of response categories in rating scales, it is expected that the change in the number of response categories in SERVPERF will lead to a change in the respondents' rating scores.

For the analysis of the data using pretest-posttest experimental designs, a variety of statistical methods can be applied. Among these methods, the most commonly used ones are Analysis of Variance (ANOVA) on the gain scores, ANOVA on residual scores, repeated measures ANOVA and Analysis of Covariance (ANCOVA)[42]. The choice for the best approach has created a debate in literature as each of these approaches hold both certain advantages and disadvantages[43]. ANCOVA is generally applied in pretest-posttest experimental designs with control groups to investigate whether there is a statistically significant difference between the posttest measures of the two groups. The use of ANCOVA reduces the within group error variance[44]. Besides, it helps with the elimination of confounds as it removes the bias of these "uncontrolled" variables that vary systematically with the experiment manipulation and allows the researcher to assess more accurately the effect of the independent variable[45]. Considering its advantage of

[41] V.Doğan, B.Y.Özkara, C.Yılmaz and Ö.Torlak, (2014). An Examination of the Optimal Number of Response Categories in terms of Data Characteristics and Data Quality: An Inference Regarding the Optimal Number of Response Categories). In the Proceedings of the 19th Annual Turkish National Marketing Congress, Gaziantep, TURKEY.

[42] D.M.Dimitrov and P.D.Rumrill, Jr.(2003). Pretest-Posttest Designs and Measurement of Change. *Work*, 20, p.160.

[43] S.B.Morris. (2008). Estimating Effect Sizes from Pretest – Posttest – Control Group Designs. *Organizational Research Methods*, 11(2), p.366.

[44] D.M.Dimitrov and P.D.Rumrill, Jr. (2003), p.161.; A. Field, 2012, *Discovering Statistics Using IBM SPSS Statistics*. Fourth Edition, London: Sage Publications, p.364.

[45] A. Field. (2012), p.364.

determining the group differences by taking into account individual differences on the covariate measure[46], for this study ANCOVA was used.

As mentioned in the literature part previously, the covariate in many studies is set up to be an indication of each participant's status on the dependent variable at the beginning of the experiment. In this study, in order to free the experiment from any potentially biasing effects of individual factors, the pretest scores of the participants are included as a covariate in the gathering and analysis of the data.

Before carrying out the main ANCOVA analysis evaluations of normality, homogeneity of variance, linearity and homogeneity of regression slopes were made in order to meet the assumptions needed prior to conducting ANCOVA[47]. The Kolmogorov-Smirnov test results show that the data is distributed normally (sig=0.457, p>0.05). The results of the Levene's test were found to be significant (p=0.115, F=2,518, p>0.05) and the variance ratio for the data was calculated as 1.49 which is below 2 indicating that the group variances are equal. The assumption of homogeneity of regression slopes between the dependent variable and the covariate assumption was not violated (sig=0.478, p>0.05) and the assumption of linearity of relationship between the pre- and post-treatment measures was met. In order to conduct ANCOVA, it is recommended that the independent variable should not affect the covariate variable[48]. In other words, the covariate and the treatment effects should be independent. In this study as the data on the covariate variable is collected before the treatment is applied, this assumption is clearly met.

Table 3
ANCOVA Test Results

Research Group	**n**	**Mean**	**Adjusted Mean**	**Standard Error**
Experiment	78	4.0454	4.050	.071
Control	73	4.3153	4.311	.073
Source	**Type III Sum of Squares**	**Mean Square**	**F**	**Sig.**
Pretest Scores	27.424	27.424	70.078	.000**
Scale Type	0.480	0.480	1.226	.270

*p≤.05; ** p≤.01

[46] L.M.Sallot ve L.J.Lyon. (2003). Investigating Effects of Tolerance – Intolerence of Ambiguity and the Teaching of Public Relations Writing: A Quasi-Experiment. *Journalism & Mass Communication Educator*, 58(3), p.268.

[47] W.K.Alford, J.M.Malouff and K.S.Osland. (2005). Written Emotional Expression as a Coping Method in Child Protective Services Officers. *International Journal of Stress Management*, 12(2), p.182-183.

[48] S.W. Huck. (2008), p.393.

As may be seen from Table 3, the pretest scores of the groups that are included in the analysis as the covariate are found to be significantly predicting the dependent variable [$F_{(}1, 148_{)}=70.08$; $p<0.01$]. The ANCOVA analysis revealed no statistically significant difference between the control and the experimental groups' evaluations of SERVPERF using 5- and 7- point Likert-type scales when the effects of the covariance are controlled [$F_{(}1, 148_{)}=1.23$; $p>0.05$]. In other words, there is no significant effect of the number of response categories in the scale on the participants' evaluations of SERVPERF after controlling for the effects of the potential confounding factors *(pretest score*s). **Thus H2 is rejected.**

In order to examine whether any differences occur for the dimensional structure of SERVPERF for the 5- and 7-point versions, Exploratory Factor Analysis (EFA) was conducted using Principal Component Analysis and Varimax rotation method. The results show that for all four groups, EFA revealed different factor structures [(EG: pre-test: 7 factors; post-test: 6 factors); (CG: pre-test: 5 factors; post-test: 7 factors)]. The highest total variance explained was achieved for the post treatment measurement of the control group (72.086%) whereas the lowest was achieved pre-test of the control group (65.081%). The results reveal that the dimensional structure of SERVPERF shows differences for both the experimental and control groups and pre- and post-treatments. As the dimensional structure of SERVPERF show differences for all of the cases **H3 is accepted**.
Regarding the results of EFA, it could be seen that the factorial structure of all of the groups show differences (varying from 5 factor results to 7 factor results). Keeping in mind that the original SERVPERF model comprises of 5 dimensions and in order to be able to make comparisons, before proceeding to CFA, for all groups the factorial structures were anchored to the original 5 factors.
To be able to assess which version of SERVPERF (5 point or 7 point) scale better fits the data, both the pre-treatment and post-treatment results of the experimental and the control groups were subjected to CFA in Lisrel 8.72 using Maximum Likelihood Estimation[49].

[49] K.G.Jöreskog and D.Sörbom. (1993). Lisrel 8: Structural Equation Modeling with Simplis Command Language. Scientific Software International.

Table 4
Pre- and Post- Treatment Goodness of Fit Statistics of SERVPERF for Experimental and Control Groups

	EXPERIMENTAL GROUP		CONTROL GROUP	
	PRE-TREATMENT	POST-TREATMENT	PRE-TREATMENT	POST-TREATMENT
χ^2 (df)	275.02 (179)	274.79 (179)	277.42 (179)	321.23 (179)
χ^2 / (df)	1.54	1.54	1.55	1.79
sig (p value)	0.00	0.00	0.00	0.00
RMSEA	0.083	0.083	0.087	0.105
NFI	0.60	0.66	0.62	0.53
CFI	0.77	0.83	0.78	0.67
SRMR	0.10	0.094	0.11	0.11
GFI	0.75	0.75	0.73	0.70
AGFI	0.67	0.67	0.65	0.62

As seen from Table 4, when the four models are evaluated and compared by examining the values of the goodness-of-fit indexes, the best fit is achieved for the post-treatment application of the SERVPERF for the experimental group where a 7-point Likert-type scale was used[50]. The results of CFA provides evidence that using a 7-point response category version of SERVPERF shows a better fit when compared to the 5-point type, taking into consideration the sample used in this study. **Thus H4 is accepted**.

Conclusion

This study is based on a pretest-posttest control group experimental design to test whether the differently pointed (i.e. 5-point vs. 7-point) response categories used in a scale lead to any statistical differences in the data characteristics regarding internal structure, normality; dimensional structure of SERVPERF; and goodness of fit analysis. Using the 5-point and 7-point versions of SERVPERF, although no statistically significant differences were spotted in terms of alpha and test-retest coefficients; a minor increase in the reliability scores was achieved as the response categories in the scale increased. According to the results of the study, the increase in the number of response categories did not lead to any changes in internal structures and normality scores. Additionally ANCOVA resulted in no statistically significant differences taking into consideration the varying number of response categories (5- and 7-point scales) on the participants' evaluations of SERVPERF.

The dimensional structure of SERVPERF showed differences between the 5- and 7- point scale formats. Also, the goodness-of-fit indexes regarding the data showed that the best fit was achieved for the 7-point scale version of SERVPERF.

[50] N.K.Malhotra. (2010), p.732-733.

In summary, the results of the analyses show that for the sample used in this study, although some minor differences were spotted in data characteristics, no statistically significant differences existed regarding the 5- and 7-point versions of SERVPERF considering internal structure, normality and differences in means. Additionally, although the dimensional structures of the scale and goodness of fit values show differences for the 5- and 7-point versions of SERVPERF these differences are only minor and hence do not provide adequate evidence that significant differences exist between the 5- and 7-point scales.

Limitations and Future Research

As with all studies, this study also has some limitations. This study has only focused on Likert-type items. Future research might also examine the effects of the number of response categories in other scale formats such as semantic differentials. The findings reported in this research are based on the results of ratings of the service quality of a certain brand of fast food restaurant (Burger King). Further research may be conducted for other fast food brands and in other settings. Besides, the findings can be further extended by using measurement instruments other than SERVPERF and by investigating participant related differences based on cultural and demographic characteristics. Furthermore, in this study only two formats of Likert-type scales (5- and 7-points) are included. Although 5-point and 7-point scale formats are "by far the most common"[51], other scale formats are also used. Therefore, scales with varying number of response categories may be added in future research. Another limitation is the sample that is made up of only students. The study may be repeated on samples other than students. In addition to the number of response categories, the effects of different levels of labeling may also be examined in future research. Although the goodness of fit statistics regarding the models achieved for the two varying scale formats provide evidence that the 7-point format shows better fit of data, further analysis may be conducted to further validate the strength of the 7-point scales as compared to 5-point scales. While this study has some limitations and the results cannot be generalized concerning the effects of the number of response categories, the findings of this study may provide implications for further research on the subject and to practitioners.

REFERENCES

Alford, W.K., Malouff, J.M. & Osland, K.S. (2005). Written Emotional Expression as a Coping Method in Child Protective Services Officers. *International Journal of Stress Management*, 12(2), 182-183.

Babakus, E. & Boller, G.W. (1992). An Emprical Assessment of the SERVQUAL Scale. *Journal of Business Research*, 24(3), 253-268. doi: 10.1016/0148-2963(92)90022-4

Bearden, W.O., Netmeyer, R.G. & Mobley, M. (1993). *Handbook of Marketing Scales: Multi-item Measures for Marketing and Consumer Behavior Research*. Newbury Park, CA: Sage.

[51] A.M.Colman, C.E.Norris and C.C. Preston. (1997), p.361.

Bendig, A.W. (1953). The Reliability of Self-ratings as a Function of the Amount of Verbal Anchoring and the Number of Categories on The Scale. *The Journal of Applied Psychology*, 37, 38-41. doi: 10.1037/h0055647

Bendig, A.W. (1954). Reliability and The Number of Rating Scale Categories. *The Journal of Applied Psychology*, 38, 38-40.

Brady, M.K., Cronin, J.J.Jr. & Brand, R.R. (2002). Performance-only Measurement of Service Quality: A Replication and Extension. *Journal of Business Research*, 55(1), 17-31.

Brown, G.; Wilding, R.E. & Coulter, R.L. (1991). Customer Evaluation of Retail Salespeople Using the SOCO Scale: A Replication Extension and Application. *Journal of the Academy of Marketing Science*, 9, 347-351.

Carillat, F.A., Jaramillo, F. & Mulki, J.P. (2007). The Validity of the SERVQUAL and SERVPERF Scales: A Meta-analytical View of 17 Years of Research Across Five Continents. *International Journal of Service Industry Management*, 18(5), 472-490. doi: 10.1108/09564230710826250

Chang, L. (1994). A Psychometric Evaluation of Four-point and Six-point Likert-type Scales in Relation to Reliability and Validity. *Applied Psychological Measurement*, 18, 205-215.

Choudhury, S. & Bhattacharjee, D. (2014). Optimal Number of Scale Points in Likert Type Scales for Quantifying Compulsive Buying Behaviour. *Asian Journal of Management Research*, 4(3), 431-440.

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, 31-36.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences*. Second Edition, New York, NY: Academic Press.

Colman, A.M., Norris, C.E. & Preston, C.C. (1997). Comparing Rating Scales of Different Lengths: Equivalence of Scores from 5-point and 7-point Scales. *Psychological Reports*, 80, 355-362. doi: 10.2466/pr0.1997.80.2.355

Cortina, J.M. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, 78, 98-104. doi: 10.1037/0021-9010.78.1.98

Cox, E.P. (1980). The Optimal Number of Response Alternatives for a Scale: A Review. *Journal of Marketing Research,* 17, 407-422.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York, NY: Holt, Rinehart & Winston.

Cronin, J.J.Jr. & Taylor, A.S.(1992). Measuring Service Quality: A Reexamination and an Extension. *Journal of Marketing*, 56(3), 243-253.

Dawes, J. (2008). Do Data Characteristics Change According to the Number of Scale Points Used?, *International Journal of Market Research*, 50(1), 61 – 77. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.417.9488&rep=rep1&type=pdf

Dimitrov, D.M. & Rumrill, Jr., P.D. (2003). Pretest-Posttest Designs and Measurement of Change. *Work*, 20, 159-165. http://iospress.metapress.com/content/7x9hgpq885t2yttq/

Doğan, V., Özkara, B.Y., Yılmaz, C. and Torlak, Ö. (2014). Katılım Düzeyi Seçenek Sayısının Veri Karakteristiği ve Veri Kalitesi Kapsamında İncelenmesi: Optimal Katılım Düzeyi Seçenek Sayısına İlişkin Bir Çıkarım (An Examination of the Optimal Number of Response Categories in terms of Data Characteristics and Data Quality: An Inference Regarding the Optimal Number of Response Categories). In the Proceedings of the 19th Annual Turkish National Marketing Congress, Gaziantep, TURKEY.

Field, A. (2012). *Discovering Statistics Using IBM SPSS Statistics*. Fourth Edition, London: Sage Publications.

Finn, R.H. (1972). Effects of Some Variations in Rating Scale Characteristics on the Means and Reliabilities of Ratings. *Educational and Psychological Measurement*, 32(7), 255-265.

Garner, W.R. (1960). Rating Scales, Discriminability and Information Transmission. *Psychological Review*, 67,343-352.

Green, J.A. & Rao, V.R. (1970). Rating Scales and Information Recovery: How Many Scales and Response Categories to Use? *Journal of Marketing*, 34, 33-39.

Howell, D.C. (1992). *Statistical Methods for Psychology*. Boston, MA: Duxbury Press.

Huck, S.W. (2008). *Reading Statistics and Research*. Fifth Edition, Boston, MA: Pearson Education, Inc.

Jain, S.K. & Gupta, G. (2004). Measuring Service Quality: SERVQUAL vs SERVPERF Scales. *The Journal for Decision Makers*, 29(2), 25-37. http://www.vikalpa.com/pdf/articles/2004/2004_apr_jun_25_37.pdf

Janssens, W., Wijnen, K. Pelsmacker, P.D. & Van Kenhove, P. (2008). *Marketing Research with SPSS*. London: Pearson Education Limited.

Jones, R.R. (1968). Differences in Response Consistency and Subjects' Preferences for Three Personality Inventory Response Formats. In Proceedings of the *76th Annual Convention of the American Psychological Association*, 247-248.

Jöreskog, K.G. and Sörbom, D. (1993). Lisrel 8: Structural Equation Modeling with Simplis Command Language. Scientific Software International.

Lai, M., Li, Yongjian & Liu, Y. (2010). Determining the Optimal Scale Width for a Rating Scale Using an Integrated Discrimination Fuction. *Measurement*, 43, 1458-1471. doi: 10.1016/j.measurement.2010.08.012

Leung, S.O. (2011). A Comparison of Psychometric Properties and Normality in 4-,5-,6 and 11-Point Likert Scales. *Journal of Social Service Research*, 37, 412-421. doi:10.1080/01488376.2011.580697

Loken, B., Pirie, P., Virnig, K.A., Hinkle, R.L. & Salmon, C.T. (1987). The Use of 0-10 Scales in Telephone Surveys. *Journal of the Market Research Society*, 29(3), 353-362.

Lozano, L.M., Garcia-Cueto, E. & Muniz, J. (2008). Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales. *Methodology*, 4(2), 73-79. doi: 10.1027/1614-2241.4.2.73

Malhotra, N. K. (2010). *Marketing Research: An Applied Orientation*. Sixth Edition, Boston, MA: Pearson.

Marlow, L., Inman, D. & Shwery, C. (2005). To What Extent are Literacy Initiatives Being Supported: Important Questions for Administrators. *Reading Improvement*, 42(3), 179. http://eric.ed.gov/?id=EJ725388

Matell, M.S. & Jacoby, J. (1971). Is There an Optimal Number of Alternatives for Likert Scale Items? Study 1: Reliability and Validity. *Educational and Psychological Measurement*, 31, 657-674. http://psycnet.apa.org/journals/apl/56/6/506/

Morris, S.B. (2008). Estimating Effect Sizes from Pretest – Posttest – Control Group Designs. *Organizational Research Methods*, 11(2), 364-386.

Oaster, T.R.F. (1989). Number of Alternatives per Choice Point and Stability of Likert-type Scales. *Perceptual and Motor Scales*, 68, 549-550. doi: 10.2466/pms.1989.68.2.549

Osteras, 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. doi: http://www.hqlo.com/content/6/1/14

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, 1-15. doi: 10.1016/S0001-6918(99)00050-5

Qin, H., Prybutok, V.R. & Zhao, Q. (2010). Perceived Service Quality in Fast-food Restaurants: Empirical Evidince from China. *International Journal of Quality & Reliability Management*, 27(4) , 424-437. doi: 10.1108/02656711011035129

Ramsay, J.O. (1973). The Effect of Number of Categories in Rating Scales on Precision of Estimation of Scale Values. *Psychometrika*, 38, 513-533. doi: 10.1177/014662168500900103

Sallot, L.M. & Lyon, L.J. (2003). Investigating Effects of Tolerance – Intolerence of Ambiguity and the Teaching of Public Relations Writing: A Quasi-Experiment. *Journalism & Mass Communication Educator*, 58(3), 251-272. doi: 10.1177/107769580305800304

Symonds, P.M. (1924). On the Loss of Reliability in Ratings Due to Coarseness of the Scale. *Journal of Experimental Psychology*, 7, 456-461. doi: 10.1177/014662168500900103

Viswanathan, M., Sudman, S. & Johnson, M. (2004). Maximum versus Meaningful Discrimination in Scale Response: Implications for Validity of Measurement of Consumer Perceptions About Products. *Journal of Business Research*, 57, 108-124. doi: 10.1016/S0148-2963(01)00296-X

Weathers, D., Sharma, S. & Niedrich, R.W. (2005). The Impact of the Number of Scale Points, Dispositional Factors and the Status Quo Decision Heuristic on Scale Reliability and Response Accuracy. *Journal of Business Research*, 58, 1516-1524. doi: 10.1016/j.jbusres.2004.08.002

Weijters, B., Cabooter, E. & Schillewaert, N. (2010). The Effect of Rating Scale Format on Response Styles: The Number of Response Categories and Response Category Labels. *International Journal of Research in Marketing*, 27, 236-247. doi: 10.1016/j.ijresmar.2010.02.004

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. doi: 10.1177/0013164404268674

Woodruff, D.J. & Feldt, L.S. (1986). Tests for Equality of Several Alpha Coefficients When Their Sample Estimates are Dependent. *Psychometrika*, 51, 393-413. http://link.springer.com/article/10.1007/BF02294063

Zhou, L. (2004). A Dimension-specific Analysis of Performance-Only Measurement of Service Quality and Satisfaction in China's Retail Banking. *The Journal of Services Marketing*, 18(6/7), 534-546. doi: 10.1108/08876040410561866