RedFame

International Journal of Social Science Studies
Vol. 11, No. 1; January 2023
ISSN 2324-8033 E-ISSN 2324-8041
Published by Redfame Publishing
URL: http://ijsss.redfame.com

Likert Items: Should(n't) We Really Care?

Raymond Doe, Bryan A. Landrum, Kaelyn M. Lewis, Matthew E. Glenn, Jacob D. Smith

[1] Lamar University, Beaumont, TX, USA

Correspondence: Raymond Doe, Department of Psychology, Lamar University, P.O. Box 10036, Beaumont TX, 77710. USA.

Received: October 21, 2022 Accepted: December 1, 2022 Available online: December 12, 2022

doi:10.11114/ijsss.v11i1.5747 URL: https://doi.org/10.11114/ijsss.v11i1.5747

Abstract

One of the controversial methodological topics in the social and behavioral sciences is the (ab)use of Likert Scale items, Likert-type items and ranked ordered response categories. The debate is whether parametric tests can be legitimately conducted on technically ordinal response categories that are represented with numbers. Participants answered survey questions on moral disengagement, where we changed the intervals of seven response categories and tested whether assigning numbers made any difference in two separate studies. The results showed that participants' ratings were not significantly different with or without numbers. Participants tend to covertly superimpose numbers where none were provided. Also, there were no significant interactions between assignment of numbers and 'intervalness'. However, ratings were significantly different between two key interval groups. Knowing the assumptions of respondents to these Likert items even without numbers could inform researchers especially if parametric tests are to be conducted.

Keywords: response categories, Likert items, numbers, intervalness, psychological distance

1. Introduction

1.1 Overview of Likert Items

Likert scales and its variations are common data collection tools in the social and behavioral sciences as well as in business and management (Alexandrov, 2010). Introduced in 1932 by Rensis Likert, it was used to measure attitudes as a latent construct (Likert, 1932). Likert scales have been misunderstood, misused and abused (see Jamieson, 2004, and Pornel & Saldaña, 2013). Some of the name variations identified in the literature are Likert-type, Likert item, and rank-ordered response categories. It is not uncommon to come across a critique of Likert items. Overall, it seems fair to say that data about a concept is collected in discrete form but transformed and interpreted as continuous data. Some of the current debate focuses on the categories of responses, adding of values/numbers, size of the scale, directionality of the scale, ordinal or interval nature of the data, and the appropriate statistical analysis of such data.

Joshi, Kale, Chandel, and Pal (2015) explored what number of points (either five or seven) on a Likert scale is preferable and the analysis of the scale. When it comes to analyzing the item responses, knowing whether the response variable is considered interval or ordinal is very important and that depends on how the instrument itself is constructed (Joshi et al., 2015). The use of parametric or non-parametric analysis seems to be the pivotal if not the consequential arguments. One school of thought considers the Likert scale ordinal while the other camp sees it as interval. Those in the ordinal school of thought (purists) believe that since the scale does not show quantitative magnitude or equal distance between any two responses, it cannot be treated as interval (Joshi et al., 2015). Sometimes the goal of the researcher is to obtain a composite score for each individual, and the distance between one individual's composite score and another can be considered interval estimates. This is how the interval school of thought (pragmatists) would interpret Likert scales. There are some researchers that create a new category for these items and refer to them as ordinal-interval scales.

Preston and Colman (2000) conducted a study examining how the number of response categories on a scale affected its reliability, validity, and discriminating power. Respondents were given a questionnaire with 11 sections about the service of a restaurant/store, each with five questions, where the only difference was the number of response options. To measure the respondents' preferences, they were also asked to complete a 101-point scale that asked about each scale's ease of use, quickness, and whether it allowed them to accurately express their feelings on the issue (Preston & Colman, 2000). They found that scales with more response categories had much higher reliability, validity, and discriminating

power compared to those with two to four options. Respondents also favored the scales with 10-points the highest, followed by the seven-point scale. However, test-retest reliability decreased for scales with ten or more response options. DeJonge, Veenhoven, and Arends (2014) also examined the number of response options on a scale, as well as the wording of those response options. They determined that both the number of options and the wording does matter, because the interval length of a response option is dependent upon the surrounding response option's level of intensity (DeJonge et al., 2014).

Another issue that may affect the psychometrics of a Likert scale is combining regular/positive items with reversed items. Using combined scales started with the goal of reducing response biases, like acquiescence bias (Suarez-Alvarez et al., 2018). One research study used a repeated measures design and had participants complete a positive, reverse, and combined form of a self-efficacy test (Suarez-Alvarez et al., 2018). The researchers found that when positive and reversed items are used in the same test, the reliability becomes flawed and the unidimensionality of the test is ruined (Suarez-Alvarez et al., 2018). It seems that although combining the type of items may help with response biases, things like response variability, test precision, discriminatory power, and reliability may suffer (Suarez-Alvarez et al., 2018).

Roster, Lucianetti, and Albaum (2015) explored the idea of using slider versus traditional radio-button scales and whether the two formats would yield significantly different results. The argument in favor of sliders is that they are less repetitious, more engaging for respondents, and the data collected using them is equal or superior to Likert scales using radio-buttons (Roster et al., 2015). Participants were randomly assigned to either the traditional radio-button or slider group and were given three sets of questions pertaining to topic sensitivity, personal importance, and likelihood of participation. There was no strong evidence for any of the arguments supporting the usage of slider scales and no statistically significant difference between the two formats (Roster et al., 2015). The two formats did not differ much regarding response rates, completion time, differences in raw mean scores, or the likelihood a respondent would participate in a survey about randomly selected topics (Roster et al., 2015).

1.2 Scales of Measurement

1.2.1 Nominal Scales

The nominal scale of measurement deals with assigning numerals to categorize objects or groups. These numerals can be labels, numbers, words, or letters (Stevens, 1946). Furthermore, Stevens (1946) proposed that one can assign nominals in two ways, (1) by the numbering of people in a group to represent these individuals. An example of this is numbering football players to represent everyone. (2) Numbering of types of groups, such as designating the number 1 to represent a female population or designating the number 2 to represent a male population. To create a nominal scale, one must have an operation for determining equality. Creating this scale will allow one to check for the number of cases, calculate mode, and calculate contingency correlation (Stevens, 1946).

1.2.2 Ordinal Scales

To determine the order of rank, such as intelligence, personality traits, or grades one would use the ordinal scale of measurement. (Stevens, 1946) reports that this scale is the most widely and effectively used by psychologist. For one to develop this scale, one must have an operation for determining greater or less. Once this is established, statistics can be used such as median and sometimes percentiles (Stevens, 1946).

1.2.3 Interval Scales

The interval scale of measurement is a quantitative scale that have equal intervals or differences between each score**.** Furthermore, interval scales do not have a true zero point, where zero represents the absence of the construct being measured. A good representation of an interval scale is a Fahrenheit Scale. Stevens (1946) proposed that for an interval scale to be created it must have both nominal and ordinal empirical operations, which are determination of equality and of greater or lesser. Furthermore, interval scales must have an operation for determining equality of intervals. This scale is used to compute mean, standard deviation, rank order correlation, and product moment correlation (Stevens, 1946).

1.2.4 Ratio Scales.

The ratio level of measurement deals with the ratios between the aspects of objects that are equal (Stevens, 1946). Unlike the interval level of measurement, ordinal scales have an absolute zero. An example of a ratio scale is the Kelvin Scale. When this scale reaches zero, it signifies that the thermal motion stops. The ratio scale of measurement can be used only when all four empirical operations exist, which are determination of equality, of greater or less, of equality intervals, and of equality of ratios. Furthermore, the ratio scale can measure fundamentals, such as length, weight, and time (Stevens, 1946).

1.3 Debate and Current Study

There is a conflict within the scientific community on how Likert-type scales can and should be used. Likert scales use

responses that are defined as ordinal data (Jamieson, 2004; Kuzon, Urbanchek, & McCabe, 1996). For example, it cannot be determined if there is an equal space between "fair" and "good," as "fair" and "good" cannot be mathematically defined, and can be categorized differently per individual (Kuzon, Urbanchek, & McCabe, 1996). However, researchers still assign numerical values to the categories within the scale and perform tests (e.g., Pearson r) on the data (Sullivan & Artino, 2013). For example, Kuzon, Urbanchek, and McCabe state that a rating of "'fair and a half'" on a medical procedure cannot be statistically valid, as there is no true numerical value behind it (Kuzon, Urbanchek, & McCabe, 1996).

Although using parametric tests on ordinal data violates assumptions, there may be no harm done in doing so. When using small data sets specifically; results using both parametric and non-parametric tests are similar when computing ordinal data (Norman, 2010). Norman compared results of both parametric and non-parametric tests on a single set of data multiple times and found that there was little difference between the two (Norman, 2010). This implies that even though using parametric tests on ordinal data may not comply with assumptions and rules regarding statistical analyses, they are still statistically sound when measuring data (Carifio & Perla, 2007). This study therefore aims to investigate how participants typically answer Likert items. Specifically, two experiments were designed to determine:

a) Whether there are differences in participants responses when they are presented with different intervals and whether adding numbers make any difference at all (Experiment 1)

b) Whether participants who have been trained on scales of measurement would respond differently compared to participants who received no training (Experiment 2).

2. Method

2.1 Participants (Experiment 1)

The study was conducted at a public university located in Southeast Texas in the United States. One hundred and nighty-two undergraduate students ranging from 18 to 51 years old participated in this study. The sample includes freshmen, sophomores, juniors, and seniors. Regarding work status, majority of the participants worked part time (see Table 1). The participants received research credits for participation in this study and this study received the Institutional Review Board approval.

Table 1. Demographic Characteristics of Participants

Characteristics (Exp. 1)	*n*	%
Gender		
Male	47	24.5
Female	143	74.5
Other	2	1.0
Classification in College		
Freshman	70	36.5
Sophomore	67	34.9
Junior	43	22.4
Senior	12	6.3
Work Status		
Full-time	24	12.5
Part-time	89	46.4
Not working	79	41.1

Note. N = 192. Participants were on average 20.63 years old (SD = 4.27).

2.2 Materials

The Moral Disengagement Scale (Bandura et al., 1996) was used in this study. The Moral Disengagement Scale (Bandura et al., 1996) consisted of 32 Likert-scale items that ranges from strongly disagree to strongly agree. In previous studies, the scale has a reliability of .82, which implies that the measure has strong internal consistency. The following is an example of one of the items: It is alright to lie to keep your friends out of trouble.

2.3 Procedures

Upon arrival at the research lab, the researchers administered an informed consent to the participants. The participants were randomly assigned to one of eight groups. The participants then answered the questions where the researchers manipulated the intervals of seven response categories (strongly disagree to strongly agree) and tested whether or not assigning numbers made any difference (see figure 1). The researchers instructed participants to mark/select anywhere on the line with an (X). A ruler was used to precisely measure the accurate marker (X) used by participants to select their response options.

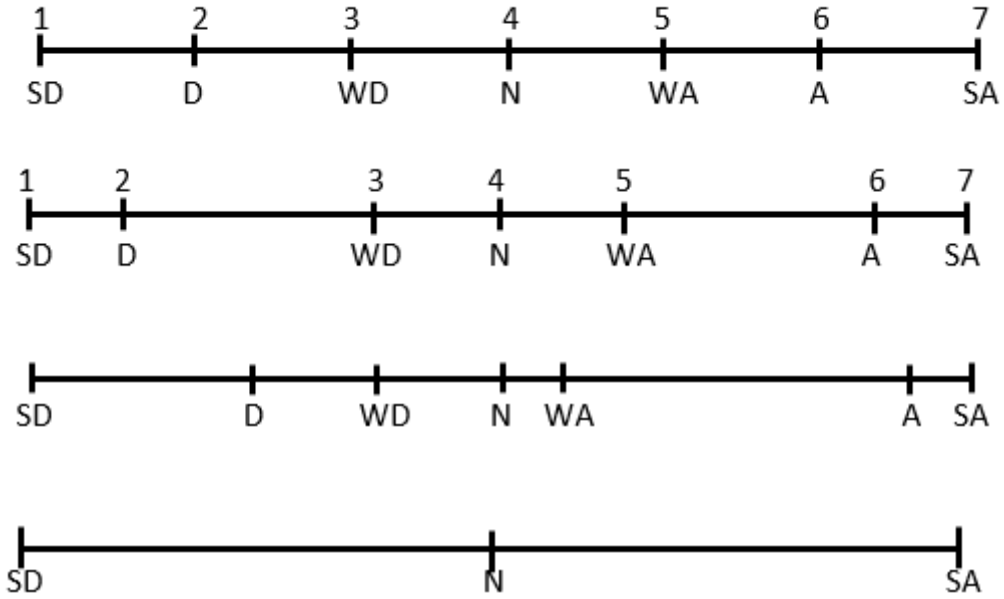

Figure 1. Sample rating scale with or without numbers

Note. Length of line = 9.5cm. SD = Strongly Disagree; D=Disagree; WD=Somewhat Disagree; N= Neutral; WA=Somewhat Agree; A = Agree; SA=Strongly Agree. In consecutive order: Equal interval, Unequal interval 1, Unequal interval 2, and Personalized interval.

3. Results

A 4x2 factorial ANOVA analysis showed that intervalness had a significant effect on moral disengagement scores, F (3,184) = 3.553, p =.016, η^2 = .054. There was no significant effect of numbers alone, F (1,184) = .235, p = .629, η^2 = .001, nor was there an interaction effect of intervalness and numbers, F (3,184) = .890, p = .448, η^2 = .014, see Table 2.

Table 2. Test of Intervalness and Values on Moral Disengagement Scores

Source	*SS*	*df*	*MS*	*F*
INTERVALNESS	8.789	3	2.930	3.553*
NUMBERS	.194	1	.194	.235
INTERVALNESS * VALUES	2.200	3	.733	.890
Error	151.718	184	.825	
Total	162.901	191		

*$p < .05$

Post hoc comparisons using Tukey's HSD test showed a significant difference between the Personalized Interval group (M = 2.167, SD = 0.89) and the Unequal Interval 2 group (M = 2.684, SD = 0.84), see Figure 2. No other group comparisons were significant.

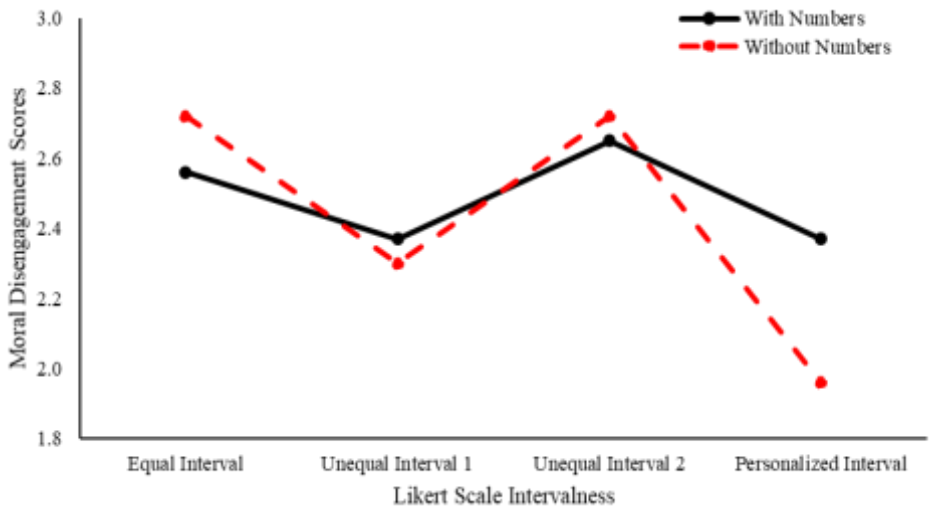

Figure 2. Group differences on moral disengagement responses

3.1 Experiment 2

Experiment 2 was similar to experiment 1 except the following:

One hundred and forty-eight participants (N =148) were randomly assigned in a 2 (Trained vs No Training) X 2 (Numbers vs No Numbers) factorial design. Based on the results of the Experiment 1, only equal interval rating scales were used for Experiment 2. The training was a video instruction on scales of measurement with graded exercises. Only participants that were proficient on the topic were allowed to proceed in the trained group of the study.

3.2 Results of Experiment 2

There were no significant differences between the trained group and the no-training group, F (1, 144) = 3.540, p = .062, η^2 = 024. Participants in the Trained group however had lower Moral disengagement scores (M = 78.297, SD = 18.477) than those in the no training group (M = 84.391, SD =20.824). As expected, there were also no significant differences between adding numbers to the scales and having no numbers on the scale, F (1, 144) = .088, p = .767, η^2 = .001. There was also no significant interaction effect. These findings supplemented the findings in Experiment 1.

4. Discussion

To find out how participants process the distances between the response options presented and the meaning attached to numbers, two experiments were conducted in this study. The findings, however, reiterate that participants do respond to questions with some pre-conceived assumptions and beliefs that are mostly ignored by researchers. For example, the results showed that not adding numbers to the response categories does not erase the psychological image of counting from left to right among participants. Participants 'saw' and impose numbers where 'none' was provided. Should that be a concern? Definitely. If parametric statistical analyses use numbers, and participants are using numbers to respond to questions anyway, it is prudent to avoid leaving the decision of what the numbers (of the response categories) represent to participants. In other words, it is best practice to impose numbers on the response options than letting the participants make that choice, nevertheless. Using Likert scales or any of its variants without numbers is just contributing to the problem and not the solution. This argument goes for the intervals as well. One, Likert item may technically be ordinal. Take the common 5-point Likert-type item from strongly disagree to strongly agree, it is difficult to argue that "disagree and a half" is possible. In the absence of a better measurement, however, when you have multiple items, and you use numbers to arrive at a mean of 2.5, you are closer to understanding that whatever construct you are measuring is definitely not in the positive range of agreement.

In addition, the findings in this study indicate that participants differed on the construct being measured when they were presented with different extreme interval options. This difference was such that participants who were offered less options (anchors and a middle point) tended to pay more attention and process their response options deeply since the image they perceive had only few options. This phenomenon has been described as end-of-scale effect (Lantz, 2013). Although participants were instructed to select any option by marking any place on the line, they tended to select the default (anchors) given. Participants usually start with some idea of what their response to an item is, but unknowingly, participants seem to be forced to change their true scores to the closest favorable option. Furthermore, the multiple response options that are within participant's favorable zone may undergo further internal processing (a concept worth investigating).

There are several studies suggesting that participants do perceive Likert-type items as non-equidistant (Bendixen & Sandler, 1995; Lee & Souter, 2010, Mundy & Dickinson, 2004; Kennedy, Riquier, & Sharp, 1996). Kuzon Jr et al. (1996) for instance has labelled the use of parametric analysis on these non-equidistant categories as one of the seven deadly sins researchers have been daring to commit. This perceived non-equidistance fades away with larger response options, symmetrical anchors, use of multiple items, and imposing numbers on the response anchors. Lubke and Muthen (2004) did argue that factor analysis works just well with Likert scale data. Norman (2010) on commenting on this issue indicated that if these numbers are reasonably distributed, inferences about means can be made. In addition, scholars have affirmed that journal reviewers seem to be making too much racket about using parametric test on Likert-type scales, since these tests are robust and can handle skewness and non-normality of data. Furthermore, the use of computers and online data collection tools look promising in solving the 'intervalness' issue for example the use of Variable–Interval Slider (VIS) (Ladd, 2009). Training participants on the technicalities seem to produce little difference (though not statistically significant).

4.1 Conclusions and Recommendations

This study set out to investigate the assumptions of users in understanding psychological distance and numbers when responding to questions. This study however only assessed part of a larger problem from the perspectives of participants. Secondly, we used typical undergraduate students in a developed country. Comparative studies on this topic as well as other unique populations such as fulltime workers and the elderly may shed more light on the topic of psychological distance. Pending further studies, researchers should use Likert scales – a combination of multiple items measuring the same construct. A seven or ten response categories with numbers should be preferred than five. Researchers should explicitly impose 'intervalness' on the response options by indicating to participants that the intervals are approximately equal. The use of numbers is preferable than not using numbers at all. Prior analysis about normality and skewness of the data should be checked and conducting a comparable nonparametric test is laudable. Use Likert scales items as a means to an end and do care about the policies that could result from the use of these self-report measures. If possible, practitioners should use non-reactive behavioral measures.

References

Alexandrov, A. (2010). Characteristics of single-item measures in Likert scale format. *The Electronic Journal of Business Research Methods, 8*(1), 1-12.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology, 71*, 364-374. https://doi.org/10.1037/0022-3514.71.2.364

Bendixen, M. T., & Sandler, M. (1995). Converting verbal scales to interval scales using correspondence analysis. *Management Dynamics: contemporary Research, 4*(1), 31-49.

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. *Journal of Social Sciences*, *3*, 106-116. https://doi.org/10.3844/jssp.2007.106.116

DeJonge, T., Veenhoven, R., & Arends, L. (2014). 'Very happy' is not always equally happy on the meaning of verbal response options in survey questions. *Journal of Happiness Studies, 16,* 77-101. https://doi.org/10.1007/s10902-013-9497-9

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education, 38,* 1217-1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Joshi, A., Kale, S, Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology, 7,* 396-403. https://doi.org/10.9734/BJAST/2015/14975

Kennedy, R., Riquier, C., & Sharp, B. (1996). Practical Applications of Correspondence Analysis to Categorical Data in Market Research. *Journal of Targeting, Measurement and Analysis for Marketing*, *5*(1), 56-70.

Kuzon, W., Urbanchek, M., & McCabe, S. (1996). The seven deadly sins of statistical analysis. *Annals of Plastic Surgery, 37,* 265-272. https://doi.org/10.1097/00000637-199609000-00006

Ladd, D. A. (2009). Everybody Likes Likert: Using a Variable-Interval Slider to Collect Interval-Level Individual Options. *ICIS 2009 Proceedings*. 100. https://aisel.aisnet.org/icis2009/100

Lantz, B. (2013). Equidistance of Likert-Type Scales and Validation of Inferential Methods Using Experiments and Simulations. *Electronic Journal of Business Research Methods, 11*(1), 16-28.

Lee, J. A., & Souter, G. N. (2010). Is Schwartz's value survey and interval, and does it really matter? *Journal of Cross*

Cultural Psychology, 41, 76-86. https://doi.org/10.1177/0022022109348920

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 1- 55.

Lubke, G. H., & Muthen, B. O. (2004). Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons. *Structural Equation Modeling, 11*, 514-534. https://doi.org/10.1207/s15328007sem1104_2

Mundy, J. & Dickinson D. (2004). Factors affecting the uptake of voluntary HIV/AIDS counselling and testing (VCT) services in the workplace. *In: HIV/AIDS in the Workplace Research Symposium*. University of Witwatersrand, June 2004, 175-193.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Science Education, 15*. https://doi.org/10.1007/s10459-010-9222-y

Pornel, J. B., & Saldaña, G. A. (2013). Four Common Misuses of the Likert Scale. *Philippine Journal of Social Sciences and Humanities, 18*(2), 12-19.

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. https://doi.org/10.1016/S0001-6918(99)00050-5

Roster, C. A., Lucianetti, L., & Albaum, G. (2015). Exploring slider vs. categorical response formats in web-based surveys. *Journal of Research Practice, 11*, 1-19.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680. https://doi.org/10.1126/science.103.2684.677

Suarez-Alvarez, J., Pedrosa, I., Lozano, L. M., Garcia-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema, 30,* 149-158. https://doi:10.7334/psicothema2018.33

Sullivan, G., & Artino, R. (2013). Analyzing and interpreting data from likert-type scales. *Journal of Graduate Medical Education, 5*, 541-542. https://doi.org/10.4300/JGME-5-4-18