Use and Misuse of the Likert Item Responses and Other Ordinal Measures

PHILLIP A. BISHOP‡[1], and ROBERT L. HERRON[1,2]†

[1]The University of Alabama, Department of Kinesiology, Exercise Physiology Laboratory, Tuscaloosa, AL, USA; [2]Auburn University at Montgomery, Department of Kinesiology, Human Performance Lab, Montgomery, AL, USA

‡Denotes professional author, †Denotes graduate student author

ABSTRACT

International Journal of Exercise Science 8(3): 297-302, 2015. Likert, Likert-type, and ordinal-scale responses are very popular psychometric item scoring schemes for attempting to quantify people's opinions, interests, or perceived efficacy of an intervention and are used extensively in Physical Education and Exercise Science research. However, these numbered measures are generally considered ordinal and violate some statistical assumptions needed to evaluate them as normally distributed, parametric data. This is an issue because parametric statistics are generally perceived as being more statistically powerful than non-parametric statistics. To avoid possible misinterpretation, care must be taken in analyzing these types of data. The use of visual analog scales may be equally efficacious and provide somewhat better data for analysis with parametric statistics.

KEY WORDS: Visual analog scale, subjective, statistical analysis, exercise science

INTRODUCTION

Likert, and Likert-type, responses are popular psychometric item scoring schemes for attempting to quantify people's opinions on different issues. The Likert scale originated with Rensis Likert (21), and has a long history of use in Kinesiology research (13, 14, 24).

The long-running issue with Likert-type scales and ordinal responses is the appropriate statistical treatment of these data. If the data are ordinal, then non-parametric statistics are typically considered the most appropriate option for analysis. If the data are interval, then parametric statistics can be used. This includes not only Likert-type scales but also other ordinal measures such as the rating of perceived exertion (RPE). For example, investigators have published research on the Rating of Perceived Exertion (3, 4), with almost all treating these data as interval rather than ordinal (1, 2, 10-12, 15). Whereas the classic Likert-scale items had 5 possible responses, the RPE scale as 14 choices (3) and the modified RPE has 10 (4).

This is an issue because parametric statistics are generally perceived as being more statistically powerful than non-parametric statistics. Knapp argues that this is not the case, regardless of perception (19).

However, the simplicity of non-parametric tests (e.g., the signed-ranks test), biases some to assign a higher status to parametric analyses than to non-parametric. Most importantly, the goal of research is to produce valid results useful for advancing the field, and valid statistical conclusions require valid statistical analyses. The purpose of this Research Note is to review current thinking on the treatment of data generated from Likert-type, and other ordinal responses and provide evidence for using alternatives.

Critiques of Likert-type Responses

In a Likert-response item with choices varying from "Strongly Disagree" to "Disagree to "Neutral", to "Agree" to "Strongly Agree", it would appear to be in the mind of the research participant whether or not there is an equal distance between each of these choices (9). Note that the above response options are "balanced" in that the items to the left of "Neutral" have an equal number of counterparts to the right of "Neutral". If the response choice is unbalanced to either side, the possibility of that item being an interval measurement seems greatly diminished.

With RPE, there is no issue of "balance", but there remains the question of the consistency of the interval between RPE ratings. For example we might expect respondents to be very sensitive to the change between "Rest" (RPE = 6) and "Fairly Light" exercise (RPE = 11) to be a larger difference than the difference between "Hard" (RPE = 15) and "Maximum" (RPE = 20) (4).

Knapp gives a useful illustration of the potential problems of Likert responses which could also be applied to RPE responses. If a response has choices, "Strongly Disagree", "Disagree", "Neutral", "Agree", and "Strongly Agree", Knapp suggests that these could readily be assigned numerical values of 1, 2, 3, 4, 5, as is often done. Knapp further argues that other numbers could be assigned such as 1, 3, 5, 7, 9, or any other linear transformation, and this would not impact the data or its analysis. In fact, Knapp points out, any ordered non-linear numerical assignment, 3, 11, 17, 23, 31 could also be made and preserves the ordinal nature of the data; however, this latter non-linear choice would have an impact on group means and whether or not parametric statistics should be used (19).

But, as Knapp illustrates, if the terms "never, seldom, occasionally, always" were used, the two middle values could be argued as being very similar, with perhaps much less distance between "seldom" and "occasionally" than between "never" and "seldom", or between "occasionally" and "always". Knapp even suggests that some would argue the two middle terms should be reordered (19). With RPE, there is less ambiguity, but it is likely that the lower parts of the scale are further apart than the upper parts of the scale, especially for those less experienced with very hard exertion.

Kuzon et al. (20) made the observation that no investigator would express the mean of a Likert-response item as "Strongly Agree and a half". But, after these descriptors are converted to numbers, investigators are comfortable doing just that; in fact the results might be (improperly) expressed as "Strongly Agree.523".

Clason and Dormoody (7) offer another critique of Likert response analyses. They suggest the following possibility for the means of a coded 5-item Likert-type response to a series of Questions:

Question Response	**1**	**2**	**3**	**4**	**5**
Group 1 (% of responses)	20	20	20	20	20
Group 2 (% of responses)	50	0	0	0	50

Regardless of group size, the mean for the two groups will be identically equal to 3, yet the two responses are obviously quite different with large difference in variance. However, it is noteworthy that this same issue could arise regardless of the type of measurement if information about the variance is not reported.

It has been long acknowledged that the extremes of a Likert-type response tend to get less use than the more central choices causing an "anchor effect" (16). Therefore, the intervals near the extremes may be further apart, than those near the center. This, by itself, disqualifies a Likert-type response as interval.

Support of Likert Responses as Interval data

Carifio and Perla (5, 6) are among the strongest supporters for treating Likert-type responses as interval data, going so far as to suggest that the Likert-responses approximate ratio data. They do make the important distinction between "Likert Scales" compared to the answers to individual questions using Likert-type responses. In their view, all true scales must necessarily include multiple-questions on a given topic whose summative score reflects the scale or measurement, and contend that a minimum of six items is necessary to create a reliable scale that measures some construct. Any particular item comprising this scale can have a response format which might or might not be a Likert-type response.

Carifio and Perla (5, 6) also argue that much of the criticism of "Likert Scales" confuses the response format from the actual multi-component measurement (i.e. Likert scale). In their view the individual items in a "scale" are not independent and autonomous, but rather must be connected in such a way as to yield a single unified result. This unified result (scale) will be more reliable and reflect the underlying construct better than will any individual item. They make the useful explanatory observation that a Likert scale need not use Likert-type responses to its individual questions, but could use a visual analog response (VAR)(5, 6). Consequently, Carifio and Perla (5, 6) make a strong argument against the statistical or interpretive analysis of individual responses, suggesting that the summative assessment of a series of items is the proper item of analysis and that such a summative assessment yields interval or ratio data. Surprisingly, Carifio and Perla (5, 6) also tout Vickers (25) as having made a strong case for the advantages of the Likert-type response assessment even though the Vickers study only used a one-item survey of pain, and not a proper "scale" by their definition given above (5, 6). Of course research measures of exertion or comfort, etc. are typically one-question measures and analyzed individually, so the six-or-more-item requirement is violated (5, 6).

Vickers (25) noted greater reliability of the Likert response compared to VAS. However, it is noteworthy that any measurement with only 5 or 7 possible discrete answers will in all likelihood, score better reliability than a measurement with 100 possible answers on a continuous measurement, i.e., if a scale or an individual item had only a single choice, it would be perfectly reliable. In a similar fashion, Vickers (25) reported that the Likert-type response to their single question of pain yielded a higher mean value than the same question posed to the same group using a VAS, and concluded that this meant that the Likert-type response was "a more responsive measure". This conclusion seems baffling when there was no criterion measurement (23).

DISCUSSION

Despite their strong support for Likert-scales (as opposed to individual Likert-type item, or, in the kinesiology case, other unequal-interval response), Carifio and Perla concede that Pearson correlations and statistical derivatives (multiple regression, factor analysis, multivariate ANOVA, and discriminant analysis) are not very tolerant of uses of ordinal data, whereas F-tests generally are robust with regard to ordinal data (5, 6, 19). Regardless of where one stands on the use of F-tests of Likert–scales or other non-equal interval measures, in any situation in which Pearson correlation-based analyses are planned, then using a VAR, or other alternative, seems to be a more conservative approach with no clear reason for not using such a scale.

In the end, it seems the most important thing to keep in mind, is that statistical analyses are not an end in themselves, but rather a means to an end. Statistics are a tool to enable investigators to think about the data, and ultimately, the population. Statistics are not a substitute for thinking about what data truly mean, and what data are showing about the population.

Along these lines, Hopkins (17, 18) is known for insisting that effect sizes be presented along with p-values. This approach does raise our awareness of Type I and Type II statistical errors. For example, when studying elite athletes, sample sizes may be small, but small effects may have great practical significance for this population, but the probability of making a Type II error is large. Conversely in situations with very large sample sizes, statistical power can be so high that impractically small changes (effects) are statistically significant but not of meaningful (practical) importance.

It seems indefensible to offer an unbalanced Likert scaled item, or any other single-measurement item as an interval measure, especially when other measurement options are available. Whether or not a balanced scale is viewed as an interval scale, alternatives to the Likert scaled, and similar items are available. Some investigators have abandoned the Likert-type response in favor of a simple visual analog scale (VAS). The VAS typically has descriptive anchors only at the two extremes, although there has not been any published research on VAS with multiple anchors.

When sample sizes are small, the participants can physically mark a 100mm line with appropriate anchors at either end.

The participant is free to mark the scale at any point desired resulting in a continuous interval measurement with scores constrained between 0 and 100, though certainly longer scales can be used. The scale can be scored by manually measuring the participant's chosen mark from the left end. A modified measure of perceived exertion using a VAS could be developed with verbal anchors only on the two extremes.

One objection to the use of VAS responses is the challenges of doing this on computerized questionnaires. This obstacle has been removed. For computerized surveys or other instruments, Reips and Funke (22) recommend their website, http://www.vas.com/, which generates VAS usable on the computer. They also offer information on the precision of these scales along with others (8, 22). This should alleviate some of the issues of large scale computerized measurements.

Despite that many psychometricists insist the data are interval (5, 6, 25) this can hardly be considered a conservative approach. Again, if Pearson correlation or analyses of variance are planned, then Likert-type or other non-interval responses should not be used. Given the recent innovations in VAR responses, there seems little reason to use Likert-type, or other non-interval responses in most research applications (22).

REFERENCES

1. Allen IE, Seaman CA. Likert Scales and Data Analyses. Quality Progress 40(7):64-65, 2007.

2. Armstrong RL. The Midpoint on a Five-point Likert-type Scale. Perceptual Motor Skills 64(2): 359-362, 1987.

3. Borg G. Physical Performance and Perceived Exertion. [Akademisk avhandling]. Lund: Lund.; 1962.

4. Borg G. Borg's Perceived Exertion and Pain Scales. Champaign, IL: Human Kinetics; 1998.

5. Carifio J, Perla R. Resolving the 50-year Debate Around using and Misusing Likert Scales. Med Educ 42(12): 1150-1152, 2008.

6. Carifio J, Perla RJ. Ten Common Misunderstandings, Misconceptions, Persistent Myths and urban legends about Likert scales and Likert response Formats and their Antidotes. J Social Sci 3(3): 106, 2007.

7. Clason DL, Dormody TJ. Analyzing Data Measured by Individual Likert-type Items. J Agricultural Ed 35: 4, 1994.

8. Couper MP, Tourangeau R, Conrad FG, Singer E. Evaluating the Effectiveness of Visual Analog Scales A Web Experiment. Social Sci Computer Rev 24(2): 227-245, 2006.

9. Dawes J. Do Data Characteristics Change According to the Number of Scale Points Used? An experiment Using 5 point, 7 point and 10 Point Scales. Int J Market Res 51(1), 2008.

10. Dunbar CC, Robertson RJ, Baun R et al. The Validity of Regulating Exercise Intensity by Ratings of Perceived Exertion. Med Sci Sports Exerc 24(1): 94-99, 1992.

11. Glass SC, Knowlton RG, Becque MD. Accuracy of RPE from Graded Exercise to Establish Exercise Training Intensity. In: Williams & Wilkins, Baltimore, MD 1992.

12. Green J, Crews T, Bosak A, Peveler W. Overall and Differentiated Ratings of Perceived Exertion at the Respiratory Compensation Threshold: Effects of Gender and Mode. Eur J Appl Physiol 89(5): 445-450, 2003.

13. Green J, McLester J, Crews T, Wickwire P, Pritchett R, Redden A. RPE-lactate Dissociation During Extended Cycling. Eur J Appl Physiol 94(1-2): 145-150, 2005.

14. Green JM, Crews T, Bosak A, Peveler W. Overall and Differentiated Ratings of Perceived Exertion at the Respiratory Compensation Threshold: Effects of Gender and Mode. Eur J Appl Physiol 89(5): 445-450, 2003.

15. Green JM, McLester JR, Crews TR, Wickwire PJ, Pritchett RC, Lomax RG. RPE Association with Lactate and Heart Rate during High-intensity Interval Cycling. Med Sci Sports Exerc 38(1): 167, 2006.

16. Guilford JP. Psychometric Methods. New York : McGraw-Hill,1954.

17. Hopkins WG. Measures of Reliability in Sports Medicine and Science. Sports Med 30(1): 1-15, 2000.

18. Hopkins WG. A Scale of Magnitudes for Effect Statistics. A new view of statistics. 2002.

19. Knapp TR. Treating ordinal scales as interval scales: an attempt to resolve the controversy. Nursing Res 39(2): 121-123, 1990.

20. Kuzon WM, Urbanchek MG, McCabe S. The Seven Deadly Sins of Statistical Analysis. Annals Plastic Surgery 37(3): 265-272, 1996.

21. Likert R. A Technique for the Measurement of Attitudes. New York, Archives of Psychology. 1932.

22. Reips U-D, Funke F. Interval-level Measurement with Visual Analogue Scales in Internet-based Research: VAS Generator. Behav Res Methods 40(3): 699-704, 2008.

23. Thomas JR, Nelson JK, Thomas KT. A Generalized Rank-Order Method for Nonparametric Analysis of Data From Exercise Science: A Tutorial. Res Quarterly Exerc Sport 70(1): 11-23, 1999.

24. Utter AC, Robertson RJ, Green JM, Suminski RR, McAnulty SR, Nieman D. Validation of the Adult OMNI Scale of Perceived Exertion for Walking/Running Exercise. Med Sci Sports Exerc 36: 1776-1780, 2004.

25. Vickers AJ. Comparison of an Ordinal and a Continuous Outcome Measure of Muscle Soreness. Int J Technology Assess Health Care 15(04): 709-716, 1999.