Analyzing Ordinal Scales in Studies of Virtual Environments: Likert or Lump It!

Abstract

Likert scaled data, which are frequently collected in studies of interaction in virtual environments, demand specialized statistical tools for analysis. The routine use of statistical methods appropriate for continuous data in this context can lead to significant inferential flaws. Likert scaled data are ordinal rather than interval scaled and need to be analyzed using rank based statistical procedures that are widely available. Likert scores are "lumpy" in the sense that they cluster around a small number of fixed values. This lumpiness is made worse by the tendency for subjects to cluster towards either the middle or the extremes of the scale. We suggest an ad hoc method to deal with such data which can involve a further lumping of the results followed by the application of nonparametric statistics.

Averaging Likert scores over several different survey questions, which is sometimes done in studies of interaction in virtual environments, results in a different sort of lumpiness. The lumped variables which are obtained in this manner can be quite murky and should be used with great caution, if at all, particularly if the number of questions over which such averaging is carried out is small.

1 Introduction

Pre–test and post–test questionnaires are part and parcel of designed experiments in human-computer interaction. Often questionnaires include questions which elicit highly subjective responses. Generally, these responses can be ordered from one extreme (such as "strongly disagree") to another (such as "strongly agree") and it is a common practice to code responses to these questions as whole numbers. These scales are known as Likert scales, introduced by Rensis Likert in 1932 in the context of psychometric analysis (Likert, 1932).

Once an experiment has been completed and the numbers are ripe for analysis, it is tempting to apply powerful statistical methods to Likert data. In fact, it is quite misleading to apply tools such as *t*-tests, ANOVA and regression to Likert scaled data as they are fundamentally rank ordered rather than interval scaled data. Although the numbers used in Likert scales are whole numbers—and hence form an equi-spaced sequence—the human responses on which they are based can be astonishingly nonlinear because subjects' interpretations of phrases such as "strongly agree" can vary widely. Indeed, in some cases the human responses can be so subjective that even the rank ordering of Likert scales is questionable—"often" to one respondent may well correspond to "sometimes" for another, and even "rarely" for another. On the other hand, statistical tools like ANOVA and regression assume that numerical scales on which data are measured behave in a regular way—for example, that 3 is as far from 4 as 6 is from 7.

Even leaving the issue of assumed regularity aside, Likert scales are also inherently "lumpy," forcing responses to a small number of choices, whereas the standard tools of inference assume that responses are measured on a continuous, interval scale. Consideration of some questionnaires leads to the conclusion that the

Henry J. Gardner*
Department of Computer Science, FEIT
College of Engineering and Computer Science
Australian National University
Canberra, ACT 0200 Australia

Michael A. Martin
School of Finance and Applied Statistics
College of Business and Economics
Australian National University
Canberra, ACT 0200 Australia

*Correspondence to Henry.Gardner@anu.edu.au

Presence, *Vol. 16, No. 4, August 2007,* 439–446

data might be even more inherently lumpy than a scale of 1 to 7 would imply. This lumpiness of the data also cautions that it should not be subject to a canned statistical analysis.

We will illustrate these points with reference to Garau, Slater, Pertaub, and Razzaque (2005). This paper contains a detailed description of a fascinating experiment where participants interacted with virtual people in an immersive virtual environment. Much of this work has been meticulously executed. But the paper contains many errors in the way that Likert data has been gathered and analyzed: survey questions have been posed which bias responses across Likert categories, averaged measures are obtained by aggregating the Likert results of different questions together, and linear regression is performed on rank ordered Likert data. These errors are, by no means, isolated to this particular paper. We hope that highlighting these issues will be of use to researchers who are concerned with the serious study of interaction in virtual environments. We do not seek to provide comprehensive advice as to how Likert-scaled data might be analyzed, but, rather, to tell a cautionary tale and offer some general guidelines. Almost every corner of the social sciences literature contains articles on how survey questionnaire data in general, and Likert scaled data in particular, may be analyzed. Two excellent general books on the design and analysis of survey questionnaires, including the analysis of Likert data, are Presser et al. (2004), and Groves et al. (2004); see also Fowler (1995) for further discussion of survey questionnaire design, and Spector (1992) for discussion of summated ratings scales, of which the Likert scale is the most well-known example.

2 The Lumpiness of Likert Results

Questionnaires need to be carefully designed if their results are to be converted to numbers and manipulated. If sufficient care is not taken, it can be easy to "force" the data to be even lumpier than a 7 point scale suggests. For example, question 1.1 of the "copresence" set of questions of Garau et al. (2005) reads:

> "During the course of the experience, did you have a sense that you were in the room with other people or did you have a sense of being alone? (With other people = 1, Alone = 7)"

This is an either/or question—it should elicit a binary response. Subjects are asked to choose between one alternative and the other—yet, oddly, a 7 point scale is retained. Even allowing for arbitrariness in respondents' behavior, it is likely that answers will be clustered strongly towards either 1 or 7.

Another example is question 1.3 of the same set:

> "To what extent did you have a sense of being in the same space as the characters? (Not at all = 1, Very much = 7)"

This is a much better posed question. A range of responses is anticipated. But perhaps the labeling of response 7 could be improved: "Very much" can mean different things to different people. One might anticipate a clustering of responses about "very much" which included people who felt that the scale should go higher than 7, say to include a category like "Completely" (see Figure 1 for a relevant cartoon).

There are other similar examples from Garau et al. (2005). But a more interesting aspect to consider is whether Likert data is naturally lumpy, even more so than a 7 point scale suggests: when a question is asked which requires a subjective response between two extremes it might be reasonable to expect that responses would bunch at either extreme or in the middle, effectively turning a 7 point scale into a much coarser 3 point scale.

The lumpiness of Likert scaled data has been extensively discussed in the psychology literature, as well as more broadly across the social sciences. Joreskog notes, in particular, that "ordinal variables do not have origins or units of measurement and should not be treated as though they are continuous" (Joreskog, 1994), and Joreskog and Sorbom (1996, p. 146) further declare that "means, variances and covariances of ordinal variables have no meaning. The only information we have are counts of cases in each cell of a multiway contin-

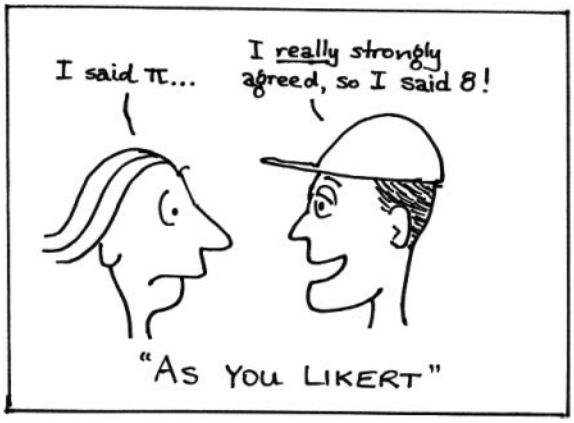

Figure 1. *This cartoon makes sense. (1 = "Strongly disagree", 7 = "Strongly agree")*

gency table"; see also Goldstein and Hersen (1984, p. 52) for similar sentiments.

The problems of how language and "the human element" affect survey responses using Likert scales has also been addressed extensively in the social sciences literature. For example, Garg (1996) reported that very different responses to a particular question could be elicited from subjects depending on whether the question was addressed using positive or negative language. Albaum (1997) noted that the Likert scale attempted to measure both the direction (agree/disagree) and intensity ("strongly" or not) of attitude, but that the scale "tends to confound the direction and intensity dimensions of attitude so there may be an under-reporting of the most intense agreement or disagreement (i.e., the extreme position of the scale)." So, for instance, he argued, "a person may hold an extreme position with little feeling, or have a middle of the road position with considerable passion; without separate questions for positions and intensity it is difficult, if ever possible, to separate these dimensions."

The human element is also important in how respondents might inadvertently promote lumpiness in Likert responses. Albaum (1997) cites three "form-related errors" resulting from subjects' psychological reactions to different item formats in questionnaires: leniency (the tendency to rate either too high or too low); central tendency (reluctance to rate at the extremes); and proximity (the tendency to rate similarly for questions occurring close to one another in the survey). The last of these errors is particularly relevant for questionnaires that group questions (for example, the question group related to "co-presence" in Garau et al. 2005), as the variation among responses is likely to be rendered artificially low by the proximity effect. Form-related errors are also discussed by Bardo, Yeager, and Klingsporn (1982), Phelps, Schmitz, and Boatright (1986), and Greenleaf (1992).

Even though many of the serious issues arising from the lumpiness of Likert data appear to be well understood in some disciplines, it is still very common for analyses of such data to fail to take into account its special structure. While the analysis of Likert-scaled data is a frequent theme in the various social science literatures, it is a sad fact that often the common wisdom or practices that guide analyses in one discipline fail to cross disciplinary boundaries.

3 Lumped Likert Results

Another sort of lumping occurs when several different Likert questions are lumped together to construct an aggregated quantity. This approach is central to some of the analysis of Garau et al. (2005) who define three aggregated quantities, "copresence," "altered participant behavior," and "perceived agent awareness." Each of these quantities is obtained by averaging the Likert scores of a small number of survey questions.

Many researchers, including us, would probably turn a blind eye to expressing the results of one Likert measurement over multiple respondents as an average (even if the median may be a better representative quantity than the average). But the averaging of several Likert measurements together to create a single score for each respondent raises more serious questions. While labeling this average score using a name like copresence is certainly evocative, there is a real issue in how one should interpret the numbers that result. Some may argue that such averaging allows the results to be interpreted as if they were continuous data drawn from a (hopefully) normal distribution. But the individual scores themselves are neither independent nor identically distrib-

uted. Indeed, one from the copresence set of questions is essentially binary, and there are only five of them in all, so there is little hope that a Central Limit Theorem would confer useful statistical properties on such an average. Having said this, it is still most likely true that higher values of the average would indicate copresence while lower values would reveal the opposite (although we have seen studies where individual Likert scaled components go in opposite directions because of the way a question is asked).

In general, the averaging of data results in its becoming smoothed or blurred so that the average is no longer restricted to the whole numbers 1 to 7. But this conferred continuity is still rather lumpy given the small number of components and little statistical comfort can be drawn from the mere act of using an average. Moreover, the variation expected in the averaged quantity can usually be expected to be less than that of most of the individual components because high and low components of the average tend to cancel each other out driving it towards the middle. This reduction in variation can have the unexpected result of covariates apparently having a more significant effect on the averaged quantity than on its individual components because a smaller effect of the independent variables is needed to cause a significant change in the averaged quantities. It does not always happen this way, though. Garau et al. (2005) spend some time discussing the results for averaged quantities and for their individual components and show that the qualitative results can be markedly different for each. The smoothing effect of averaging can often blur what is really happening so much that no effect can be detected. This is a well-known peril of aggregation of statistical data.

More serious, though, is the question of why averages have been made in the first place. In particular, if you have questions that force lumpiness into the responses that you then combine with other questions that might have a more smooth response curve, as is the case for the set of questions dealing with copresence in Garau et al. (2005), then the data so formed can become essentially meaningless. In attempting to capture a concept such as copresence, a simple average of answers to linked questions needs to be justified carefully before it can be accepted as a reliable measure of a newly defined concept. We also note the tendency for human subjects to behave somewhat arbitrarily in responding to questions of the kind posed here. Some subjects will give the same rank for all questions, while others will swing wildly. Averaging tends to smooth these effects out, the concomitant information loss making it difficult to discover consistent patterns in the original data.

Finally, let us presume that we are wrong, and that the research community wishes to use averaged quantities to streamline the processing of Likert data. If this is the case, careful consideration needs to be made as to whether the component questions should be weighted differently from one another. Why is a simple, equally weighted average the right thing to do, and does the result represent a meaningful concept? Many social researchers would prefer to use a technique like Factor Analysis or Principal Components Analysis to detect structure and form meaningful, low dimensional constructs in multidimensional data. Even then, combining such apparently lumpy measures is a risky business.

4 ANOVA, Regression and Covariates

Analysis of Variance is a powerful statistical technique that attempts to detect differences among several groups through an assessment of whether the between groups variation is large compared to within group variation. Although it assumes that the dependent measurements are based on interval scales and that they exhibit normality, it can be used for discrete measurements, such as error counts, providing the distribution of the residuals is not too far away from normal. As such, ANOVA may be a useful tool, but only if the response is not too lumpy, and if the scale on which the response is measured is an interval scale—both rather large ifs in the current context.

Many introductory statistics texts, as well as the documentation to most commonly available statistical packages such as SPSS, argue that ANOVA procedures are reasonably robust to mild, or even moderate, departures from the model assumptions, particularly the assumption that the underlying data is drawn from a normal

distribution. While this is true to an extent, ANOVA is somewhat less robust to other departures from its assumptions—in particular from its common variance, or homoscedasticity, assumption. So, what can we expect when assumptions are violated, as they certainly are by Likert scaled data? First, Likert scaled data tends to exhibit excess skewness compared with normal data, and so the significance levels reported for ANOVA-based tests are likely to be compromised as the distribution of the usual test statistic will not be close to the usual Student's *t* distribution. As a result, the powerful test (low probability of Type 2 error) promised by standard ANOVA will have a significance (probability of Type 1 error) that is different (usually higher) than what is reported. Second, the variance of a variable measured on a Likert scale is likely to be an unreliable measure of the true, underlying spread of a corresponding variable measured on an interval scale (Joreskog, 1994). As a result, the variance measure fundamental to constructing the ANOVA test is likely to be similarly unreliable, calling into question the entire basis of the ANOVA test (which relies on comparing variation explained by the model with unexplained variation).

Simple linear regression is an explicit line fitting technique that allows the prediction of changes in a response when independent variables change by a certain amount. As such, changes in the independent variables need to be explicitly quantifiable, and the scale on which these quantities are measured needs to be interpretable in an unambiguous way. Central to this requirement is that a deviation of, say, 1 in an independent variable should have the same meaning regardless of start and end points (3 must be the same distance from 4 as 6 is from 7) otherwise the fitted line is not a line at all. But Likert scales are rank ordered, not interval scaled, and there is no guarantee—indeed it is very unlikely—that the true scale is linear. Simple linear regression also makes the tacit assumption that the independent variable is measured without error, or at least that such error is negligible. We simply do not think that this is the case for this type of subjective response data.

Garau et al. (2005) perform some regression of Likert results against a Likert covariate, "computer usage" ("the extent to which you use a computer in your daily activities; 1 = Not at All; 7 = Almost all the time"). This computer usage question was imprecise (Is "daily" the same as "weekly"? Does 4 correspond to using a computer for half a day every day?) and individual differences would give rise to uncertainty. A measurement of 4 for one individual might well be 3 for another, or 5 for another. The independent variates, on the horizontal axis, are, themselves, reasonably variable. This condition can significantly shift the regression line since least squares, the criterion used to fit the line, concerns itself with vertical (response) rather than horizontal (covariate) variation in the data. The scale of measurement of the covariates is also inherently nonlinear, begging the question of why a linear fit to the data could have been expected at all.

Garau et al. (2005) plot one set of regression lines in their Figure 4. Consider how the raw data points of this figure might look if error bars were attached that represented the lumpy uncertainty in the vertical and the horizontal directions. Given the likely nonlinearity of both horizontal and vertical scales, the use of linear regression to associate them is, at best, optimistic. As it is, the fitted straight line relating copresence to factor 3 ("responsive" virtual humans) increases with decreasing computer usage, and achieves the maximum possible copresence score, of 7, at a "computer usage" score of 4. Extrapolation of copresence for factor 3, for hypothetical respondents with a computer usage score of 1, would yield a predicted copresence score of around 12 on a 1 to 7 scale. One might argue that our criticism is spurious, as the true relationship may, in fact, be nonlinear. We could only respond by saying "Exactly!" as we cannot imagine that the true relationship is, indeed, linear given that the independent variate should be interpreted as ranked data rather than as interval scaled data. As Russell and Bobko (1992) attest in the title of their related article, the Likert scale is simply too coarse for us to be comfortable about such a regression analysis.

5 More Lumping Needed?

As we have said, the experiment reported by Garau et al. (2005) seems to have been carefully designed and carried out. Their major conclusions are based on

post-experiment interviews as well as statistical data and many of them are probably correct. Of the statistically based conclusions, the reported correlation of measures of social anxiety with participant behavior is probably large enough (at a reported r value of 0.55) to be a real effect even though we would not condone this analysis. Indeed, Joreskog (1994) suggests that Pearson product moment correlation coefficients commonly underestimate the correlation between ordinal variables, and so the relationship may, in fact, be stronger than the authors realized! In any event, the stated value for the correlation (0.55) is likely to be misleading in understanding the strength, if not the nature, of the relationship. Although lumped measures are shown graphically, a significant effect on experimental condition is only reported for one component question of a lumped measure—that relating to "personal contact." We do not have access to the experimental data, but we speculate that not only the reported experimental condition 3, but also condition 4 might result in a significantly higher sense of personal contact had the statistical analysis been done correctly. The reported effects of computer usage on personal contact are based on a regression analysis about which we are quite dubious and we would not support any conclusions on the basis of the reported results.

Even if the major conclusions will only change slightly, the statistical analysis of Garau et al. (2005) should be redone, and there are many accessible tools that can be used. As a starting point, Wilcoxon and Mann-Whitney tests are the rank based equivalents of paired and independent, two sample t tests, respectively, and they are just as easy to use. The Kruskal-Wallis test is the nonparametric equivalent of one way ANOVA, and the Friedman test is the nonparametric version of two way ANOVA. Other modern nonparametric tools such as permutation procedures or bootstrap algorithms are also applicable. All of these procedures are available in most modern statistics packages. They are certainly available in the *R* statistical environment, which is free and open source (www.r–project.org).

One idea that researchers might like to consider is to acknowledge that Likert scales might be lumpy and to lump them even further. If a question expects, or obtains, a near binary response, then lumping the data into two bins would make it possible to use techniques such as logistic regression or, even simpler, nonparametric procedures such as the Sign test to detect significant structure. Sometimes simpler scales such as binary responses are better than over-engineered 7 point Likert scales, as the outcomes are less prone to misinterpretation or to the arbitrariness that results when humans are required to differentiate their attitudes on too fine a scale.

If the data can be coerced on to a binary or even a ternary scale, one simple way to analyze the data is through a classification tree approach, a nonparametric classification method that works well for categorical data, particularly binary data. Classification trees, part of a set of procedures known as CART, are increasingly popular as analytical tools for analyzing categorical data, and are implemented in statistical packages such as R. The underlying idea of classification trees is to initially partition the data into subgroups homogeneous in the response according to values of the covariates, and then to recursively repeat this process until the variation in the response is adequately explained by the fitted tree model. CART is described in detail by Breiman, Friedman, Olshen, and Stone (1984).

While CART is a sophisticated statistical tool for analyzing data of this type, we anticipate that many researchers would benefit from the development of simpler tools for understanding structure in their data, particularly for exploratory purposes. So here is an idea for an ad hoc, post hoc, statistical analysis of data which is found to be too lumpy on a Likert scale of 1 to 7:

1. Look again at the question that was asked. Is it reasonable to think that the question was forcing data to be too lumpy? If so, we assume that the lumpiness is either towards either end of the scale or clustering at the middle as well as at both ends.
2. If the question qualifies as one that generates an essentially binary response, then aggregate the data into either of the two extreme bins. Distribute data points near the middle equally between the two bins, randomly assigning the final point if there is an odd number of points in the middle.

Data that aggregates into a "low, middle, high" pattern might reasonably be aggregated into three bins, and analyzed using, say, a Kruskal-Wallis nonparametric test procedure.

3. Perform pairwise comparisons across the relevant factors using either the Median test (for between subjects experiments) or the Sign test (for within subjects experiments), adjusting the significance level required of the tests according to Bonferroni's rule if there are several tests to carry out (for example, setting the significance level at $0.05/g$ if g is the number of tests to be conducted).

While this suggested method is simple and ad hoc, it would allow researchers to respond to patterns in their data that suggest the original 1 to 7 scale was too fine, and to detect broad patterns in their data without resorting to a sophisticated analysis. Of course, it may be that a sophisticated analysis will follow, but in many cases an informal technique such as the one suggested here might be sufficient to understand broad features of the data.

6 Conclusion

Although we have focused our discussion on one particular paper, the issues that we have discussed here should be relevant to a wide range of studies of interaction in virtual environments and, indeed, to any studies that rely on Likert scoring of questionnaires. There are probably a couple of reasons why researchers in human-computer interaction are reluctant to use nonparametric statistics. The first might be that many have been brought up with parametric statistics and feel more culturally comfortable with it. Indeed, for many non-statisticians, the so-called standard tests may be the only ones to which they have had a reasonable exposure. The second reason is the issue of power. The common wisdom is that parametric tests are more powerful (in the statistical sense) than nonparametric procedures. This rubric is true only provided the underlying assumptions behind the parametric methods are satisfied! Generally, powerful tests confer the ability to design experiments with not only the required significance level but also an acceptably small probability of Type 2 error with as small a sample as possible. So less powerful nonparametric tests can make your experiment more expensive by requiring larger samples to achieve reasonable levels of both Type 1 and Type 2 errors. But cheaper tests that are fundamentally flawed can be just as damaging to scientific credibility as more appropriate tests may be to the budget.

An alternative to worrying about expensive experiments is to be a little more relaxed about the p values needed to show significance. Readers might be interested to know just how arbitrary the 0.05 gold standard of significance actually is! It arose from a convenient, but arbitrary, personal opinion expressed by R. A. Fisher in 1926. The whole story is discussed in Reese (2004).

References

Albaum, G. (1997). The Likert scale revisited: An alternate version. *Journal of the Market Research Society, 39,* 331–348.

Bardo, J. W., Yeager, S. J., & Klingsporn, M. J. (1982). Preliminary assessment of format-specific central tendency and leniency error in summated rating scales. *Perceptual and Motor Skills, 54,* 227–234.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth.

Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage.

Garau, M., Slater, M., Pertaub, D.-P., & Razzaque, S. (2005). The responses of people to virtual humans in an immersive virtual environment. *Presence: Teleoperators and Virtual Environments, 14*(1), 104–116.

Garg, R. K. (1996). The influence of positive and negative working and issue involvement on responses to Likert scales in marketing research. *Journal of the Market Research Society, 38,* 235–246.

Goldstein, G., & Hersen, M. (1984). *Handbook of psychological assessment.* New York: Pergamon Press.

Greenleaf, E. (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research, 29,* 176–188.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: Wiley.

Joreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59,* 381–389.

Joreskog, K. G., & Sorbom, D. (1996). *PRELIS 2: User's reference guide*. Chicago: Scientific Software International.

Likert, R. (1932). The method of constructing an attitude scale. *Archives of Psychology, 140,* 44–53.

Phelps, L., Schmitz, C. D., & Boatright, B. (1986). The effects of halo and leniency on co-operating teacher reports using Likert-type rating scales. *Journal of Educational Research, 79,* 151–154.

Presser, S., Rothjeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J. et al. (2004) *Methods for testing and evaluating survey questionnaires*. New York: Wiley.

Reese, R. A. (2004). Does significance matter? *Significance, 1*(1), 39–40.

Russell, C. J., & Bobko, P. (1992). Moderated regression analysis and Likert scale: Too coarse for comfort. *Journal of Applied Psychology, 77,* 336–342.

Spector, P. E. (1992). *Summated rating scale construction*. Thousand Oaks, CA: Sage.