Likert scales and questions - uses and abuses

Geoffrey Darnton
Warwick Manufacturing Group (WMG), University of Warwick

gdarnton@requirementsanalytics.com

Abstract: when conducting business and management research, the most common strategy employed for collecting primary data is either interviews or questionnaires (or both). When questionnaires are used to collect primary data a very common approach used is to construct questions which can be called Likert-type questions or Likert scales. Numbers are allocated to the question responses of each question.. Often, this is followed by performing arithmetic or statistical operations on the allocated numbers. In many cases encountered by the author, the analysis has even included techniques such as parametric statistics and factor analysis. This paper explains why such simplistic approaches are completely invalid and should never be used. It goes on to explain how analysis can be done whilst avoiding typical hazards. Often, the writers of such papers do not understand or explain levels of measurement. Of course, Likert scales are at an ordinal level of measurement which would normally preclude the use of arithmetic, statistical, factor analysis techniques there is an additional problem of reliability because different people will interpret terms such as strongly disagree and disagree, differently. Likert was aware of these problems when writing the original paper in 1932 although at that time the term "levels of measurement" was not in use. This paper provides approaches and suggestions for avoiding the problems of data analysis when using Likert-type questions. This paper should be of assistance to those who intend to use Likert-type questions in the questionnaire.

Keywords: levels of measurement, Likert questions, factor analysis.

1. Introduction

This paper is inspired by the number of student dissertations the author has had to deal with in the past few years where the students collected primary data by means of questionnaires containing Likert -type questions. Typically, the analysis of the primary data has involved converting the Likert question responses into numbers and then performing arithmetic and statistical operations on the resulting numbers.

The paper gives more detail of these abuses of Likert -type questions and the analysis approaches used. It discusses and presents ways in which Likert -type questions can be used more reasonably as a precursor to data analysis. This paper also deals with the controversy in the literature about the appropriate use of Likert -type questions.

2. Common Abuses

A typical problematic dissertation proceeds like this. The overall structure of a dissertation is: front matter (any mandatory preliminary pages such as declarations), abstract, keywords, acknowledgements, contents, lists of tables and diagrams, acronyms and abbreviations) introduction (setting out background, aim, objectives, summary of approach, dissertation structure); literature review; methodology (setting out key methodological choices, methods of analysis, anonymized primary data sources, perhaps sample questions, summary comments); results; analysis; discussion (to include discussion of differences between primary and secondary data); conclusions; back matter (list of references at a minimum). Although many dissertations place the literature review before methodology, there are very good reasons why generally, methodology should come immediately after the introduction chapter. This is because there are important comments to be made in methodology about how the literature review will be conducted. For example, it is very unlikely that all possible references related to the topics discussed in the dissertation can be read and taken into account, therefore some kind of sampling is needed when selecting sources of secondary data to consider. Searches can be made of red suitable publishers of peer-reviewed journal papers using random sampling or even weight the available sources by a measure such as citation count.

Key methodological choices for such dissertations, usually include quantitative, deductive, and survey by questionnaires. Sampling is very rarely simple probabilistic sampling because it is common for the population not to be well defined or available.

Questionnaires typically contain Likert-type questions. There is usually poor or no discussion of levels of measurement. Analysis proceeds by allocating numbers to question responses, the most common encountered being: strongly disagree = 1; disagree = 2, neutral =3, agree = 4, strongly agree =5 (or the other way round). Essential matters missing are level of measurement, scales of measurement, and reliability.

Taking reliability first, if reliable questions are questions that mean the same thing different people, then the ends of the Likert scales of 1-5 are likely to be an reliable because different people will take different views as between strongly disagree and disagree or agree and strongly agree. Likert recognises this difficulty, but not presenting any discussion of levels of measurement, does present examples of five-point and three-point scale (Likert, 1932). Turning to levels of measurement, it should be clear that Likert -type scales of 1-5 or 2-4 are at an ordinal level of measurement (because it cannot be assumed that the distances between the available question responses are in any sense a quick distant). This means that primarily, facing those numbers generated from the Likert -type questions cannot be subjected to normal arithmetic or statistical operations. Verification of that only requires looking at the case where disagree = 2 and agree = 4, it can be seen immediately that if those numbers can be processed by normal arithmetic operations, the result is that agree = 2* disagree! Such a conclusion is obvious nonsense.

To compound the initial level of measurement problems associated with Likert -type questions and converting responses to numbers, the problematic dissertations usually proceed by performing factor analysis and regression calculations on the resulting numbers. This usually demonstrates a serious misunderstanding of the appropriateness of factor analysis generally, and regression given the problems of performing arithmetic and statistical operations on ordinal measurements.

Analysis usually starts by trying to use and apply some tests to establish that the use of factor analysis is appropriate. An example of such a test is Cronbach's Alpha. The author of this paper has never seen one of these dissertations discuss Cronbach's alpha or even contain a citation to a reference covering the use of such a test. One example of an appropriate citation is Cronbach (1951). Other tests that may be included in the preliminaries to factor analysis are the KMO (Kaiser–Meyer–Olkin) test (based on Kaiser, 1970 and/or Kaiser and Rice, 1974) , and Bartlett's Test of Sphericity (based on Bartlett, 1937)).

The use of factor analysis in these problematic dissertations is not usually resolved by a careful discussion of factor analysis in terms of what it is trying to do, and how to do it properly. In essence, factor analysis can help where there is a large dataset involving multiple dimensions. Initially, factor analysis looks at the dataset to see if most of the variance can be accounted for by a small number of hypothetical factors. Correlation analysis is used to see if there are any significant relationships between the factor analysis hypothetical factors and the dataset dimensions. A discussion of this is normally missing in the problematic dissertations. More fundamentally, the dissertation dataset is usually formed from ordinal scales derived from the Likert -type questions. This means that factor analysis is not an appropriate form of analysis because it really needs interval or ratio levels of measurement before it can be used. Doing so can introduce insurmountable problems of incommensurability.

What has been observed in the problematic dissertations is that very few present the results of converting Likert -type questions to numeric scales of 1-5 by means of simple bar charts for the question responses available. There is nothing wrong with doing that. Slightly more dissertations present results in terms of average scores obtained by calculating the average score for each question following converting question responses to numbers. This is not a valid arithmetic operation on ordinal scales. Most dissertations do not present questionnaire results using descriptive statistics but ignore the issue and go straight on to present the results of inappropriate tests to check whether valid use can be made of factor analysis. Of course, these tests are being used in an inappropriate context and should not be used in such a way.

3. Likert

In preparing this paper, it proved impossible to obtain a clean original copy of Likert's paper, made more difficult by the lack of responsiveness by APA.

Likert's subject of interest was the "measurement of attitudes" (Likert, 1932).He starts with a detailed discussion about "what is an attitude?". There has been considerable research over the years about attitudes and they measurement, but the author of this paper would venture to provide operational definition of attitude which will be used throughout this paper: an attitude is a predisposition to behave in a particular way in response to a particular situation. Therefore, attitude measurement forms part of disposition measurement. Some key problems include: how consistent over time is someone's predisposition to behave in a particular way?; How consistent are a person's attitudes, even in the simple sense the consistency of answers to the same question when a person is asked the question multiple times with time gaps between being questioned?; How contextually-dependent our propensities to behave, and how to measure such contexts? This paper is not going to try to answer any of such questions, but they are important and the

interested reader is referred to Likert's original paper and detailed issues of consistency and correlation addressed by writers such as Cronbach (1947, 1951).

When constructing a Likert -type questionnaire, it is important to create sets of questions that are relevant to the attitudes being measured. Likert does this when he explains "The project conceived in 1929 by Gardner Murphy and the present writer aimed first of all to present a wide array of problems having to do with these five major "attitude areas": international relations, race relations, economic conflict, political conflict and religion. The attitude areas best covered in the questionnaire are those of race relations, international re- lations, and economic conflict. It was our conviction that very high specificity would exist among the elements used in our questionnaire, except insofar as the questions clearly dealt with the same issues" (Likert, 1932; p11). Likert shows a number of ways he presents questions and allocate numbers to the responses:

4. Likert Questions and Scales

Likert uses a variety of ways to present questions and allocate codes to the available answers. These can be questions with three possible answers

14. Should the United States enter the World Court?

YES	?	NO
(4)	(3)	(2)

or set of questions with codes for the particular responses

16. How much military training should we have?

(a) We need universal compulsory military training. (1)

(b) We need Citizens Military Training Camps and Reserve Officers Training Corps, but not universal military training (2)

(c) We need some facilities for training reserve officers but not as much as at present. (3)

(d) We need only such military training as is required to maintain our regular army. (4)

(e) All military training should be abolished. (5)

or the most common which is questions and answers on a scale from strongly disagree to strongly agree and codes 1-5.

17. All men who have the opportunity should enlist in the Citizens Military Training Camps

Strongly Approve	Approve	Undecided	Disapprove	Strongly Disapprove
(1)	(2)	(3)	(4)	(5)

A quick look at the numbers suggested by Likert for the possible answers in all these three questions shows that there is no usefulness whatsoever in changing the responses to the numbers. Doing so will mislead easily anyone who attempts to make any use of the numbers in an arithmetic sense. Indeed, there is difficulty in relation to the sample question 16 above in that it is not obvious that these questions should be treated as mutually exclusive and it is certain that the codes associated with questions could not be used in any arithmetic or statistical sense. Another point can be made about the above sample questions taken from the Likert's paper: the paper is about attitude measurement. As mentioned above, an attitude is a propensity to behave in a particular way in a certain situation. Therefore, attitude and belief are not the same things. The sample questions are more concerned with measuring belief than attitude. Of course, a point can be reached were a belief is significant in forming someone's attitude.

5. Discussions about using Likert questions

In this paper is can be seen that the use of Likert -type questions and the allocation of codes to the answers is seriously problematic and should not be done without very careful attention to the details of problems that can arise. Dissertations that set their methodology to use Likert -type questions in questionnaires to collect primary data should be completely discouraged. Having said that, why do so many dissertations adopt a methodology of questionnaires with Likert -type questions followed by completely inappropriate factor analysis? A lot of responsibility must lie with the dissertation supervises themselves who do not pick up the problems discussed above, do not encourage their students to deal with the problems, or indeed, are directly responsible for recommending the approach to the students. Keep in mind that Likert was concerned with measuring attitude. Many dissertations are not concerned with problems of measuring attitudes. Questionnaires in dissertations to collect primary data are concerned to ask questions but many topics, not just attitude measurement. In most cases, it would be better to construct questionnaires using whatever questions are appropriate, then checking the reliability and validity of the questions before proceeding to use it for the primary data collection.

Given the potential problems of Likert -type questions and scales and the fact that Likert's original paper appeared as far back as 1932, there must have been subsequent debate about the use of Likert's work. Indeed, there has been and we should look at some of it.

Louangrath summarizes "Empirical evidence shows that non- Likert scale (0,1,2,3) is 92% reliable while the Likert-type of scale had 90, 89, and 88% reliability. Validity test showed that non-Likert scale was 93% reliable, while the Likert-type scale had 89, 61, and 57% precision. Through Monte Carlo simulation and NK landscape method for optimization, the ability of information retention for non-Likert scale was 0.96 and 0.73, 0.75, and 0.77 for Likert scales. We standardize the scale efficacy in a 5.0 system, the non-Likert scale is 4.73 and 2.35, 2.45, and 2.41 for Likert scales." (Louangrath, 1018; p50).

Joshi et al (2015) summarize the variety of views about Likert scales and levels of measurement thus: "The first school of researchers and statisticians consider Likert scale as ordinal scale. They argue that choices or responses are arranged in some ranking order. However, as this scale doesn't show the relative magnitude and distance between two responses quantitatively, it can't be treated as interval scale. The other school interprets this dilemma from a different perspective, stating that when the aim of the researcher is to 'combine' all the items in order to generate a 'composite' score for an individual rather than separate analysis of single item responded by all individuals, then this individualistic summative score (for all the items) of a participant shows a sensible realistic distance from the individual summative score of another individual; hence, can be labeled as 'interval estimates." Joshi et al;p300).

Louangrath, and, Joshi et al., provide summaries of different aspects of Likert scales. In this sense, they neither agree nor disagree with each other, so both contributions can be combined when constructing an overall view about Likert-type questions and Likert scales stop both their papers help the researcher as gateways into related literature that addresses "controversies about Likert" (hereinafter called "Likert controversies").

For the purposes of this paper, it is sufficient to note that in the light of the identified Likert controversies, students should be steered well away from dissertations that include the collection of primary data by the blind use of Likert-type questions.

The problematic dissertations that use the unfortunate combination of questionnaires based on Likert-type questions followed by factor analysis should be considered to be fundamentally flawed methodologically if they fail to handle the resulting Likert controversies or can demonstrate that factor analysis is appropriate by the blind use of the pre-factor analysis tests discussed above. Therefore, it dissertation based on that methodological set of choices including Likert -type questionnaires followed by factor analysis will normally be assessed as a failure. Over many years, this paper's author has seen several dissertations fail for this reason and has only seen one dissertation that achieved a pass mark because it included a detailed and careful discussion of the advantages and disadvantages of Likert questionnaires, but the pass was only a bare pass because of the failure to deal adequately with the problems related to the use of factor analysis.

6. Factor Analysis

An example of scenarios where the use of factor analysis is appropriate, can be expressed like this: there is a dataset that has been created for a particular research situation by using a number of measures of dimensions that may be considered important when trying to construct a model of what may be required to achieve a

particular result. That statement is too abstract, so let's look at a couple of real scenarios the author of this paper has looked at statistically. The first example scenario is concerned with trying to measure bank customer satisfaction. The dataset is constructed by obtaining a number of measures that may be related to customer satisfaction: statements by the customer of feelings of satisfaction or dissatisfaction; measures of the activities in the bank performed by the customer; measures of the context in which the customer interacted with the bank (such as location of the bank, bank layout, staff demeanour, ease of performing the required transaction, waiting time, behaviour of other customers, knowledge of bank staff…). The question of interest is: "what are the most important factors resulting in good customer satisfaction?". The second scenario concerns a dataset of a number of measures of factors that influence a person's perception of the tranquillity of a rural environment. The dataset includes measures such as: a person's perception of tranquillity; background noise from natural phenomena such as birds or wind; background noise from activities by other human beings such as parking, social intercourse, eating and drinking, using radios… The question of interest is what are the factors that contribute or detract from a person's feeling of tranquillity?

For both of the examples, there is a dataset containing measures of multiple dimensions such as those mentioned. Factor analysis is used initially on the whole dataset to ask a question like "how many hypothetical factors can be constructed to account for the variance in the dataset?". Factor analysis will produce an initial list of hypothetical factors and how much of the variance is accounted for by each factor. The next stage is to explore correlation between the factor analysis hypothetical factors and the dimensions in the dataset being explored. The hope is statistically significant correlations will be identified in order that the factor analysis hypothetical factors can be expressed in terms of the dimensions in the dataset.

This means that the student decides to use factor analysis of the dataset obtained by the dissertation questionnaire, needs to be sure of the identity of the specific dimensions being explored by the questionnaire.

There are several books available which discuss factor analysis and its applications, such as (Pett et al, 2003) and if the researcher is using SPSS for the factor analysis, a helpful book is (Pallant, 2020).

7. Summary

following the discussion above it is very difficult to understand why dissertations use that unfortunate combination of Likert questionnaires and factor analysis at the heart of the dissertation research methodology. A summary of all the key points made may help the person who is still not yet sure about the issues raised.

Whether the methodological decision is made by the student or the supervisor, it is very unwise to decide to make use of Likert -type questionnaires, and doing so should be considered as deprecated. Likert's original paper was published in 1932. Since that date, there has emerged a considerable amount of research literature about several key issues related to Likert-type questionnaires. This paper summarizes the literature in terms of a set of "Likert controversies". These controversies remain today and have not been resolved since 1932. *Prima facie*, Likert scales are at an ordinal level of measurement in the absence of appropriate validity tests to demonstrate that scores allocated to Likert -type question responses, form a genuine scale and the points can be treated as equidistant (in other words, they can be treated as interval scale equivalents). 5-point Likert -type questions have reliability problems at the ends of the scale (this can be mitigated to some extent by the use of 7- or 10-point questions). A Likert -type questionnaire where question responses are allocated numbers is unlikely to yield a dataset that is suitable for factor analysis unless great care has been taken to group the questions according to themes to be investigated in the research (a useful way to do this would be to include thematic analysis among the techniques to analyse it dissertations secondary data). Common tests (such as Cronbach, KMO, and Bartlett) are unlikely to be valid pre-factor analysis tests of the suitability of a dataset created from a Likert -type questionnaire by allocating codes to question responses.

8. Recommendations

Following what has been said above, the key recommendation must be: don't use as a research methodology, Likert -type questionnaires followed by factor analysis!

It is obvious that some researchers have a preference for Likert -type questionnaires. In the absence of a convincing discussion that it is reasonable to consider scales resulting from allocating scores to question responses, the resulting scales must be considered as ordinal. That can work if the researcher summarises the questionnaire results by using some device such as bar charts to represent the responses to each question individually. There should be no attempt to compute an average response for questions using such scores. This

also means that in terms key methodological choices, a research using Likert -type questionnaires cannot be quantitative or deductive. It can only be qualitative and inductive. There may be situations where Likert -type questionnaires can be used to obtain inductive results. Steer students away from collecting primary data by means of a Likert -type questionnaire, followed by factor analysis.

References

Bartlett, M. S. (1937). "Properties of sufficiency and statistical tests". *Proceedings of the Royal Statistical Society*, Series A 160, 268–282

Cronbach, Lee (1947) "Test "reliability" and its determination Psychometrika – Vol. 12 (1)

Cronbach, Lee J. (1951) "Coefficient Alpha and the internal structure of tests". Psychometrika – Vol 16(3) September, 1951 pp 297-334.

Joshi, Ankur; Kale, Saket; Chandel, Satish and Pal1, D. K. (2015) Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology 7(4): pp396-403*

Kaiser, Henry F. (1970). "A second generation little jiffy". Psychometrika. ***35*** *(4): 401–415.*

Kaiser, Henry F.; Rice, John (1974). "Little Jiffy, Mark IV". Educational and Psychological Measurement. ***34****: 111–117.*

Likert, Rensis (1932) "A Technique for the Measurement of Attitudes". Archives of Psychology, No 140

Louangrath, P. (2018). "Reliability and Validity of Survey Scales." *Inter. J. Res. Methodol. Soc. Sci.*, Vol., 4, No. 1: pp. 50-62.

Pallant, Julie (2020) SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS (7th ed.) Open University Press

Pett, Marjorie (Marg) A; Lackey, Nancy R; and Sullivan, John (2003) Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research. Sage Publications