Likert Scales

…are the meaning of life:

Dane Bertram

Note: A glossary is included near the end of this handout defining many of the terms used throughout this report.

Likert Scale \lick-urt\, n.

Definition: A psychometric response scale primarily used in questionnaires to obtain participant's preferences or degree of agreement with a statement or set of statements. Likert scales are a non-comparative scaling technique and are unidimensional (only measure a single trait) in nature. Respondents are asked to indicate their level of agreement with a given statement by way of an ordinal scale.

Variations: Most commonly seen as a 5-point scale ranging from "Strongly Disagree" on one end to "Strongly Agree" on the other with "Neither Agree nor Disagree" in the middle; however, some practitioners advocate the use of 7 and 9-point scales which add additional granularity. Sometimes a 4-point (or other even-numbered) scale is used to produce an ipsative (forced choice) measure where no indifferent option is available. Each level on the scale is assigned a numeric value or coding, usually starting at 1 and incremented by one for each level. For example:

Figure 1. Sample scale used in Likert scale questions

Origin: Named after Dr. Rensis Likert, a sociologist at the University of Michigan, who developed the technique. His original report entitled "A Technique for the Measurement of Attitudes" was published in the Archives of Psychology in 1932. His goal was to develop a means of measuring psychological attitudes in a "scientific" way. Specifically, he sought a method that would produce attitude measures that could reasonably be interpreted as measurements on a proper metric scale, in the same sense that we consider grams or degrees Celsius true measurement scales (Uebersax, 2006).

From http://www.performancezoom.com/performanceszoom_fichiers/likert.gif

Example: Suppose we are comparing the opinions of Masters and PhD students in CPSC.

Please indicate how much you agree or disagree with each of the following statements:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
1. The "U of C • This is now" website is easy to use.	1	2	3	4	5
2. The "My U of C" website is easy to use.	1	2	3	4	5
3. The "Peoplesoft Student Center" website is easy to use.	1	2	3	4	5

Analysis: Each specific question (or "item") can have its response analyzed separately, or have it summed with other related items to create a score for a group of statements. This is also why Likert scales are sometimes called summative scales. For our example we will evaluate the results as a whole using descriptive statistics, and also the specific results for question 1 (see Mann-Whitney U test section below).

Individual responses are normally treated as ***ordinal data*** because although the response levels do have relative position, we cannot presume that participants perceive the difference between adjacent levels to be equal (a requirement for *interval data*). In practice, many researchers *do* treat Likert scale response data as if it were interval data; however, from a statistical standpoint this can be dangerous. For example, there is no way to ensure that participants view the difference between "agree" and "strongly agree" the same as they might view the difference between "agree" and "neutral."

> "The average of 'fair' and 'good' is not 'fair-and-a-half'; which is true even when one assigns integers to represent 'fair' and 'good'!"
>
> – Susan Jamieson paraphrasing Kuzon Jr et al. (Jamieson, 2004)

The raw data for our example is outlined in Table 1 below. The participant responses have been grouped according to Masters and PhD students in order to help relate this data to the statistics we will calculate in the following sections.

Participant ID	Category	Q1. President	Q2. GSA	Q3. CSGS
1	MSc	4	4	3
2		3	4	3
3		4	3	2
4		2	3	4
5		5	3	3
6		4	2	2
7		3	3	3
8		4	4	4
9	PhD	3	4	3
10		2	5	2
11		2	4	2
12		4	1	3
13		1	3	2
14		2	2	3
15		4	3	3
16		1	1	2

Table 1. Raw Data

Tables 2, 3, 4, and 5 provide two variations of the descriptive statistics that can be calculated for the above data. Tables 2 and 3 show the median, mode, range, and inter-quartile range for the raw data where Table 2 treats all the responses together as a whole and Table 3 breaks down the same statistics into our two participant categories (Masters and PhD students) in order to aid in the comparison of these groups.

	Median	Mode	Range	Inter-quartile Range
Q1. U of C	3	4	4	2
Q2. My U of C	3	3	4	1.25
Q3. Peoplesoft	3	3	2	1

Table 2. Descriptive Statistics 1A

	Median		Mode		Range		Inter-quartile Range	
	MSc	PhD	MSc	PhD	MSc	PhD	MSc	PhD
Q1. U of C	4	2	4	2	3	3	1	1.5
Q2. My U of C	3	3	3	4	2	4	1	2.25
Q3. Peoplesoft	3	2.5	3	3	2	1	0.5	1

Table 3. Descriptive Statistics 1B

Tables 4 and 5 go on to aggregate the number of responses for each Likert level in each question where Table 4 again treats all the responses as a whole while Table 5 distinguishes between Masters and PhD student responses.

		Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
Q1. U of C	#	2	4	3	6	1
	%	13%	25%	19%	38%	6%
Q2. My U of C	#	2	2	6	5	1
	%	13%	13%	38%	31%	6%
Q3. Peoplesoft	#	0	6	8	2	0
	%	0%	38%	50%	13%	0%

Table 4. Descriptive Statistics 2A

		Strongly disagree		Somewhat disagree		Neither agree nor disagree		Somewhat agree		Strongly agree	
		MSc	PhD	MSc	PhD	MSc	PhD	MSc	PhD	MSc	PhD
Q1. U of C	#	0	2	1	3	2	1	4	2	1	0
	%	0%	25%	13%	38%	25%	13%	50%	25%	13%	0%
Q2. My U of C	#	0	2	1	1	4	2	3	2	0	1
	%	0%	25%	13%	13%	50%	25%	38%	25%	0%	13%
Q3. Peoplesoft	#	0	0	2	4	4	4	2	0	0	0
	%	0%	0%	25%	50%	50%	50%	25%	0%	0%	0%

Table 5. Descriptive Statistics 2B

Methods: Depending on how the Likert scale questions are treated, a number of different analysis methods can be applied:

1. Analysis methods used for individual questions (ordinal data):
   - bar charts and dot plots
     - ***not*** histograms (data is not continuous)
   - central tendency summarised by median and mode
     - ***not*** mean
   - variability summarised by range and inter-quartile range
     - ***not*** standard deviation
   - analyzed using non-parametric tests (differences between the medians of comparable groups)
     - Mann-Whitney U test (see below)
     - Wilcoxon signed-rank test
     - Kruskal-Wallis test

2. When multiple Likert question responses are summed together (interval data):
   - all questions *must* use the same Likert scale
   - must be a defendable approximation to an interval scale (i.e. coding indicates magnitude of difference between items, but there is no absolute zero point)
   - all items measure a single latent variable (i.e. a variable that is not directly observed, but rather inferred from other variables that are observed and directly measured)
   - analyzed using parametric tests
     - analysis of variance (ANOVA)

3. Analysis methods used when reduced to nominal levels of agree vs. disagree:
   - Chi-square test
   - Cochran Q test
   - McNemar test

Mann-Whitney U test:

To give an example of how you might evaluate a single Likert scale question we will use the Mann-Whitney U test (also called the Mann-Whitney-Wilcoxon, Wilcoxon rank-sum test, or Wilcoxcon-Mann-Whitney test) to compare the opinions of Masters vs. PhD students with respect to the usability of the main U of C website (question 1 from the example). This is a non-parametric test, and is therefore well suited to our Likert scale data as we cannot presume that the underlying population fits a normal distribution (or any other parameterized distribution for that matter). This test requires that our two samples be statistically independent (i.e. results from one sample do not affect results in the other sample), and that the observations be ordinal. We can use this method to test the null hypothesis that there is an equal probability that an observation from one sample will exceed an observation from the other sample—essentially stating that the two samples come from the same population.

Running the Mann-Whitney U test:

1. Calculate the U statistic.

To calculate the U statistic we combine the observation values from both samples and write them down in rank-order. Below each observation value we mark which sample it came from (alternating between the two samples when the same observation value is repeated and can be seen in both samples). This has been done with the observation values for question 1 as follows (P = PhD sample, M = MSc sample):

Rank-ordered:	1,	1,	2,	2,	2,	2,	3,	3,	3,	4,	4,	4,	4,	4,	4,	5
Origin sample:	P,	P,	M,	P,	P,	P,	M,	P,	M,	M,	P,	M,	P,	M,	M,	M

Next, moving from left to right, we take each observation from sample 1 (Masters students' responses) and count the number of observations from sample 2 (PhD students' responses) occurring after it (to the right) in the list. When there are matching responses (the same observation value) from each sample we count a half.

For example, with the first Masters student response we have the following:

			Tie													
Rank-ordered:	1,	1,	2,	2,	2,	2,	3,	3,	3,	4,	4,	4,	4,	4,	4,	5
Origin sample:	P,	P,	M,	P,	P,	P,	M,	P,	M,	M,	P,	M,	P,	M,	M,	M

Since there is a tie, we count 0.5 and then 1 for each of the PhD responses (after the tie) appearing to the right of this Masters response in the list giving us a count of 5.5.

For the second Masters student response we have the following:

							Tie									
Rank-ordered:	1,	1,	2,	2,	2,	2,	3,	3,	3,	4,	4,	4,	4,	4,	4,	5
Origin sample:	P,	P,	M,	P,	P,	P,	M,	P,	M,	M,	P,	M,	P,	M,	M,	M

Again we have a tie, so we count 0.5 and then 1 for each additional PhD response appearing to the right giving us a count of 2.5.

For the third Masters student response we don't have a tie, so we simply count 1 for each PhD response appearing to the right giving us a count of 2. This process continues until we've calculated a count for each of the Masters student responses. These counts are added together to give us the U statistic when starting the counting process with sample 1:

U_{MSc} = 5.5 + 2.5 + 2 + 1.5 + 0.5 + 0 + 0 + 0
= 12

Similarly, we perform the same calculation for each observation from sample 2.

In this example there is no tie for the first PhD student response and all of the Masters student responses come after it in the list, giving a count of 8. Just as before, we continue this process for each PhD student response yielding the following U statistic when starting with sample 2:

U_{PhD} = 8 + 8 + 7.5 + 7 + 7 + 6.5 + 4.5 + 3.5
= 52

Note: A convenient check to ensure your numbers are correct is to ensure that:
$U_1 + U_2$ = (# of observations in Sample 1) x (# of observations in Sample)

This check works because in the most extreme case, all the values from one sample would come before the values from the other sample. Thus, moving left to right, each of the counts would be either 8 or 0 in the example above.

2. After calculating the U statistics, consult the table of critical values for the Mann-Whitney U distribution (Table 6) using the lower of the two calculated U statistics (U_{MSc} = 12 in this case). Note: Table 6 is only a portion of the full table adapted from (Bissonnette, 2004).

- n_1 = # of observations in sample 1 (8 in this case)
- n_2 = # of observations in sample 2 (8 in this case)
- α = level of significance

If your U statistic is below the value indicated in the table, you can reject the null hypothesis and state with a given confidence level that the results/samples are significantly different. So, in this case we can see that our U statistic (U_{MSc} = 12) is below the value indicated in the table (13) at a significance level of .05 when working with two samples of 8 observations each. Thus, we can reject the null hypothesis at the .05 level and state that the MSc and PhD samples are significantly different in their opinion of the main U of C website.

n_2	α	n_1						
		3	4	5	6	7	8	9
3	.05	--	0	0	1	1	2	2
	.01	--	0	0	0	0	0	0
4	.05	--	0	1	2	3	4	4
	.01	--	--	0	0	0	1	1
5	.05	0	1	2	3	5	6	7
	.01	--	--	0	1	1	2	3
6	.05	1	2	3	5	6	8	10
	.01	--	0	1	2	3	4	5
7	.05	1	3	5	6	8	10	12
	.01	--	0	1	3	4	6	7
8	.05	2	4	6	8	10	13	15
	.01	--	1	2	4	6	7	9
9	.05	2	4	7	10	12	15	17
	.01	0	1	3	5	7	9	11

Table 6. Mann-Whitney U Distribution Critical Values

Likert Scale Strengths:

- simple to construct
- likely to produce a highly reliable scale
- easy to read and complete for participants

Likert Scale Weaknesses:

- central tendency bias
  - participants may avoid extreme response categories
- acquiescense bias
  - participants may agree with statements as presented in order to "please" the experimenter
- social desirability bias
  - portray themselves in a more socially favourable light rather than being honest
- lack of reproducibility
- validity may be difficult to demonstrate
  - are you measuring what you set out to measure?

Glossary:

inter-quartile range

- the difference between the 3rd quartile (Q_3) and the 1st quartile (Q_1); the middle 50% of the data
  1. Use the median to split data in two (don't include the median in either half)
  2. Lower quartile value is the median of the lower half; upper quartile value is the median of upper half

interval scale

- numbers assigned to responses indicate magnitude of difference between items, but there is no absolute zero point (i.e. differences between pairs of measurements can be meaningfully compared)

latent variable

- a variable that is not directly observed, but rather inferred from other variables that are observed and directly measured

median

- the middle number in a sorted list of data

mode

- the most frequent number in a set of data

non-comparative scaling

- each item is scaled independently from the others (ex. How do you feel about X?)
- contrasts comparative scaling where items are compared with each other (ex. Do you prefer X or Y?)

non-parametric

- underlying population does not have a pre-defined distribution (e.g. a normal distribution)

one-tailed test

- only interested in a difference in a single direction (i.e. hypothesis predicts the direction of difference ahead of time)

ordinal scale

- classification into ordered categories, but there is no information about the magnitude of differences between categories

psychometric

- measurement of psychological variables such as attitudes, abilities, personality traits, etc.

quartile

- any of three values that segment sorted data into four equal parts
  First quartile (Q_1) cuts off the lowest 25% of the data
  Second quartile (Q_2) is the same as the median
  Third quartile (Q_3) cuts off the highest 25% of the data

range

- difference between the largest and smallest value in a set of data

statistical independence

- the occurrence of one event makes it neither more nor less probable that the other event occurs

two-tailed test

- interested in the difference as well as the direction of the difference

unidimensional

- measures only a single underlying trait

Resources:

Likert scale – Wikipedia, the free encyclopedia (http://en.wikipedia.org/wiki/Likert_scale)

- good overview of the method; serves as a good jump-off page for finding out more about specific analysis methods, related scales, and background information

Likert Scale – Dr. Del Siegle's home page, Neag School of Education, University of Connecticut (http://www.gifted.uconn.edu/siegle/research/instrument%20Reliability%20and%20Validity/Likert.html)

- collection of commonly used Likert scales (categories/levels) for various types of attitude measurement (agreement, frequency, importance, quality, likelihood, etc.)

Further information on the various analysis methods mentioned above can be found as follows:
Mann-Whitney U test - http://en.wikipedia.org/wiki/Mann-Whitney_test
Wilcoxon signed-rank test - http://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test
Kruskal-Wallis test - http://en.wikipedia.org/wiki/Kruskal-Wallis_test

Analysis of Variance (ANOVA) - http://en.wikipedia.org/wiki/Anova
Chi-square test - http://en.wikipedia.org/wiki/Chi-Square
Cochran Q test - http://en.wikipedia.org/wiki/Cochran%27s_theorem
McNemar test - http://en.wikipedia.org/wiki/McNemar-Test

References:

Bissonnette, Victor L. "Statistical Tables. " Victor Bissonnette's Home Page. 23 Mar. 2004. Dept. of Psychology, Berry College. 23 Oct. 2007 <http://fsweb.berry.edu/academic/education/vbissonnette/tables/tables.html>

• contains various statistical look-up tables, specifically the one abbreviated as Table 6 in this report
• other areas of this site outline and demonstrate various statistical methods through the use of interactive applets

Jamieson, Susan. "Likert Scales: How to (Ab)Use Them." Medical Education 38 (2004): 1217-1218.

• short article outlines some common pitfalls seen in practice when using Likert scales
• specifically it elaborates on the inherent problems in treating Likert scale result data as interval data when it should generally be treated as ordinal data
• also serves as a concise summary of other work in the area expressing similar concerns

Kuzon WM. Jr, Urbanchek MB., and McCabe S. "The seven deadly sins of statistical analysis." Ann Plastic Surg 37 (1996): 265-72

• (included for completeness) the original paper that Jamieson paraphrases in her paper (referenced above) which I then quote in this report

Likert, Rensis. "A Technique for the Measurement of Attitudes." Archives of Psychology 140 (1932): 1-55.

• Dr. Likert's original publication about the scales that would later come to bear his name
• unfortunately I was unable to obtain a copy of this paper in digital format (or otherwise) due to the age of its publication

Mogey, Nora. "So You Want to Use a Likert Scale?" Learning Technology Dissemination Initiative. 25 Mar. 1999. Heriot-Watt University. 20 Oct. 2007 <http://www.icbl.hw.ac.uk/ltdi/cookbook/info_likert_scale/index.html>.

• page referenced from the Wikipedia article (see the resources section of this report)
• offers a concise, high-level overview of Likert Scales as well as the descriptive and inferential techniques that can be applied to them
• disappointingly light on specifics when it comes to examples and analysis

Page-Bucci, Hilary. "The Value of Likert Scales in Measuring Attitudes of Online Learners." HKA Designs. Feb. 2003. 20 Oct. 2007 <http://www.hkadesigns.co.uk/websites/msc/reme/likert.htm>.

• a report on the virtues of Likert scales in the context of measuring the attitudes of online learners
• outlines a brief overview and comparison of various related scales
• discusses advantages, disadvantages and some of the reliability and validity concerns

Shneiderman, Ben. Designing the User Interface: Stategies for Effective Human-Computer Interaction. 3rd ed. Addison Wesley Longman, Inc., 1998. 136-143.

• provides excerpts from a questionnaire for user interface satisfaction which shows a number of Likert scale type questions and how they can be applied to the Human-Computer Interaction field

Uebersax, John S. "Likert Scales: Dispelling the Confusion." Statistical Methods for Rater Agreement. 31 Aug. 2006. 20 Oct. 2007 <http://ourworld.compuserve.com/homepages/jsuebersax/likert.htm>.

• does a great job of defining and differentiating the various terms used in connection with Likert scales
• gives concrete examples of each variation and pointed arguments for the risks involved in the common misguided assumptions of ordinal vs. interval data