THE ROLE OF SCALES IN TEACHING STATISTICS FOR SOCIAL SCIENCE STUDENTS

Peter Sedlmeier
Chemnitz University of Technology, Germany
peter.sedlmeier@phil.tu-chemnitz.de

Statistical analysis for social scientists very often means statistical analysis of some questionnaire data. The meaning of the numbers obtained in such analyses depends very much on the kind of scales used. In this paper it is shown that the meaning of numbers can also depend on how exactly the scales are constructed. First, some background information about how scales of the same type (e.g., interval scales) can considerably differ in meaning is given and then a series of study results with interval, ordinal, and nominal scales that demonstrate these differential effects are reported. It is argued that such results can easily be replicated in statistics classes. It is further argued that due to the preponderance of scales in social science research, statistics courses should put an emphasis on teaching the correct use and interpretation of scales. Here, demonstrations such as the ones described in this paper can play a helpful role.

INTRODUCTION

A number is just a number – this is what most university students in the social sciences think before they enter their first statistics class. Then, in the course of the class they might be presented with a problem like the following, which appears in different versions in statistics textbooks: Mr. Miller (in German textbooks, he comes from the U.S.) conducted a study on the causation of traffic accidents. In particular, he was interested in which of the following four groups was most likely to cause accidents: female-Caucasian, male-Caucasian, male-colored, or female-colored. He coded the groups with “1,” “2,” “3,” “4,” respectively, calculated the mean and obtained a value of 2.1 which let him conclude that the Caucasian males are the most reckless drivers (because the mean of 2.1 is closest to the value of “2”). Usually, at least some students don’t object to that conclusion but most sense that something is not quite correct here and find for themselves the difference between what they later learn to call nominal and interval scales. Scales of different kinds are omnipresent in the social sciences and the numbers obtained from them are the basis for all kinds of statistical analyses. Often, these analyses use procedures that rely on arithmetic means and on variances and therefore require interval scales (e.g., ANOVA, *t*-test, Pearson correlation). If the numbers do not fulfill the scale requirement, the results may be hard to interpret or outright misleading as illustrated in the example above (in that example it would, of course, not even make sense to calculate a median). Therefore, teaching statistics cannot be separated from teaching the right use of scales.

This paper deals with a more subtle, but nonetheless quite relevant problem of the use of scales: How does the information contained in scales as used in questionnaires influence people’s responses and what does that mean for the results of statistical analysis? I begin with some background information on how the way scales are constructed may affect respondents’ answers. Then I report the results of some studies I did in my statistics classes, and finally I will discuss what students can learn from such exercises.

THE HIDDEN MEANING OF SCALES

Why should there be some hidden meaning in the way scales are built? According to Grice’s (1975) famous *cooperative principle*, we expect conversational contributions to be as informative as necessary, relevant, truthful, brief, orderly, and unambiguous. If now the meaning of a sentence is not fully specified, we usually draw so called *pragmatic implications* (Harris and Monaco, 1978) for which we use all kinds of relevant and available information. If the conversational contribution is an item on a questionnaire, the scale that defines the allowable responses is such an information. (See Schwarz, 1999, for an overview on the impact of the way a scale is constructed on the answers to be expected.)

An example: When asked to rate how successful they had been in life, on a 11-point scale, ranging from -5 (not at all successful) to +5 (extremely successful) 13% of respondents chose a value between -5 and 0. However, when the numeric values were changed to range from 0 (not at all successful) to 10 (extremely successful) the proportion of responses in the lower half of the scale increased to 34% (Schwarz, Knäuper, Hippler, Noelle-Neumann and Clark, 1991). Apparently, the pairing of "not at all successful" with "-5" let participants interpret it as "presence of failure" whereas the pairing of the identical label with "0" might have driven them to an interpretation of "absence of success." Note that the only difference was in the numbers used that, of course, can be easily linearly transformed into each other.

A second example: Schwarz, Hippler, Deutsch and Strack (1985) asked respondents about their daily TV consumption. For one group they provided response alternatives beginning with "up to ½ hour" to "more than 2½ hours" in increments of ½ hours. In this group, only 16.2% of the respondents indicated to watch TV more than 2 ½ hours per day. This proportion increased to 37.5% in another group, which received a scale ranging from "up to 2½ hours" to "more than 4½ hours" in increments of ½ hours. Here it seems that participants used the idea that "middle is normal" – the middle values (between the two middle categories) were 1½ hours in the first and 3½ hours in the second case – and adjusted what they remembered accordingly.

The last example involved ordinal scales, but the tendency to prefer middle values was also found for nominal scales. Both test-makers and test-takers in multiple choice tasks seem to prefer the middle options (Bar-Hillel and Attali, 2002).

RESULTS OF STUDIES IN THE STATISTICS CLASSROOM

Students are usually much more interested in data they generate themselves than in arbitrary samples. Therefore, to demonstrate the impact of scales on responses, I conducted several studies during regular statistics classes for undergraduates in social sciences and psychology. These studies involved interval, ordinal and nominal scales using study-relevant questions and judgments. In the following, I will give an example each of the results found in these studies. In all the studies, students received questionnaires to collect some self-generated data to be used in the statistics classes. The questionnaires differed only in the critical items discussed below, which were included at a later position on those questionnaires. Questionnaires were distributed from a stack that contained the two versions in alternating order. In none of the studies reported below, any of the students noticed that there were two different questionnaire versions, as confirmed by asking them afterwards.

Interval Scale

Two groups of students in a statistics class for social science students received questionnaires that contained, among several other items, the following question: "How relevant do you think this course is for your main subject of study?" One group (*unipolar version, n* = 37) was asked to give a number between *0 = irrelevant* and *100 = very relevant* and the other group (*bipolar version, n* = 38) should give a number between *-50 = irrelevant* and *50 = very relevant*. Figure 1 shows the results in the form of box plots (50 was added to the bipolar ratings to make the two scale versions comparable).

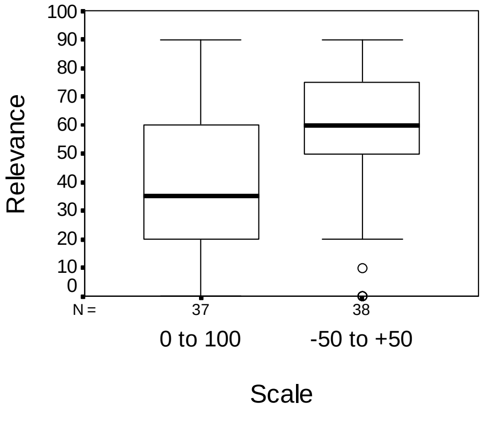

Figure 1: Boxplots showing the differences in ratings for an unipolar (left) vs. a bipolar scale (right)

Judgments made with the unipolar scale turned out to be remarkably lower than those made with the bipolar scale. A possible explanation for the result could be that participants overall had a positive opinion about the relevance of the course and therefore tended to use only the positive part of the bipolar scale whereas they used the whole range of the unipolar scale. Anyway, these results would allow to say that the relevance was judged to be markedly below 50% in the unipolar version and well above 50% in the bipolar version – a kind of judgment often used in evaluative comments. Here, however, the difference is solely due to the type of scale used.

Ordinal Scale

In the same statistics class, the two groups of students also received the following question: "What is your estimate: how much time will you spend on average to prepare for one session of this course and go over the lesson again afterwards? Please check one of the following:" In each of the two versions, five choices were offered. For the first group (*small-scale condition, n* = 37) these choices were "less than 5 minutes," "5 to 15 minutes," "16 to 30 minutes," "31 minutes up to 1 hour," and "more than 1 hour." The respective categories in the other group (*large-scale condition, n* = 36) were "less than 30 minutes," "31 minutes up to 1 hour," "1 up to 1.5 hours," 1.5 up to 2 hours," and "more than 2 hours." The results indicated that the time range offered strongly affected students' judgments: Whereas only 8.1% of the students in the small-scale version intended to spend 1 hour or more to prepare for a session of the course and review the lesson afterwards, 30% of the students in the large-scale condition did so. Apparently, students had used the information given in the scale, that is, taking the values in the middle of the scale as "normal values" and adjusted their memories towards those values.

Nominal Scale

To examine the impact of ordering response alternatives in nominal scales, the question students in a class for psychology majors found on a questionnaire was: "If you were to judge the quality of a course, which of the following four criteria would be the most important for you?" Again, students were randomly divided into two different groups. The first group (*n* = 38) received the criteria in this order (in one line): "exam orientation," "practical orientation," structuredness," and "difficulty." In the other group (*n* = 38), the ordering was "practical orientation," exam orientation," difficulty," and structuredness." Figure 2 shows that it made in general a difference whether an alternative was placed in the middle of a sequence of categories or on the outside.

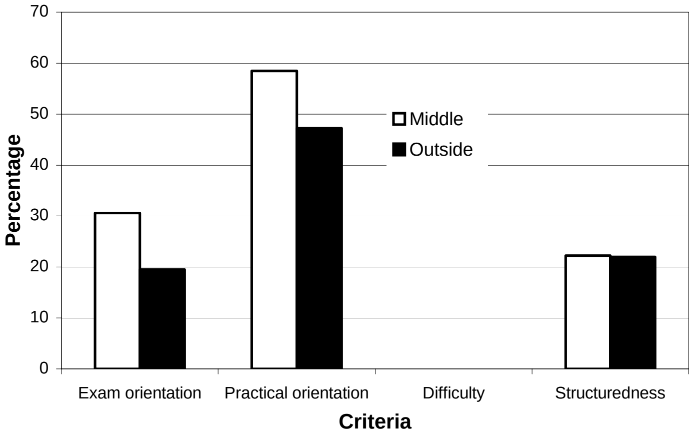

Figure 2: Percentages with which any of four evaluation criteria were chosen as "most important." Dark bars represent results for the criteria when they were placed at the beginning or end of a sequence of the four alternatives and white bars represent the respective results for middle places.

This difference is only minimal for "structuredness" but clearly present for "exam orientation" and "practical orientation" (interestingly, difficulty was not judged important despite strong evidence in evaluation studies, that it can influence evaluations substantially). Again, the result is consistent with the assumption that values in the middle are considered "normal," and therefore tend to be preferred over outside values.

DISCUSSION

The results were presented to students in a session that was dedicated to the meaning and use of scales. When seeing the differences created by seemingly small variations in labeling and ordering, students were quite puzzled, which evoked a lively discussion on the potential impact of the manipulations used, and similar ones, on the interpretation of research findings.

Each of the results reported above was replicated at least once and it seems that the effects seem to be quite stable. So studies like these can be used for demonstration purposes in statistics classes. It might, however, be difficult to do replications in the same class because students might become sensitive to slight differences in the questionnaires after the specifics of the study as well as the results are revealed to them. Building up this sensitivity is exactly the purpose of the demonstration.

It is well known that students master statistical issues much better when they are directly involved and when they learn by doing (e.g., Sedlmeier, 1999). This was the case in the studies reported above. The basic message to students that can be conveyed through such exercises is that the meaning of numbers heavily depends on their origin. This involves basic issues about different types of scales but it also involves the interpretation of numbers that have been (linearly) transformed or that stem from different scales that are to measure the same things.

Many of the social science students who visit statistics classes will use some sort of scales later on or will at least use results that stem from studies using such scales. So it is all-important that they are able to interpret the results correctly. Classroom demonstrations such as the one described here are not only fun for students but also have the potential to sensitize them to the right use and interpretation of scales.

REFERENCES

Bar-Hillel, M. and Attali, Y. (2002). Seek whence: Answer sequences and their consequences in key-balanced multiple-choice tests. *The American Statistician*, 56, 299-303.

Grice, H. P. (1975). Logic and conversation. In D. Davidson and G. Harman (Eds.). *The Logic of Grammar*, (pp. 64–75). Encino, CA: Dickenson.

Harris, R. J. and Monaco, G. E. (1978). Psychology of pragmatic implication: Information processing between the lines. *Journal of Experimental Psychology: General*, 107, 1–27.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist,* 54, 93-105.

Schwarz, N., Hippler, H. J., Deutsch, B. and Strack, F. (1985). Response categories: Effects on behavioral reports and comparative judgments. *Public Opinion Quarterly*, 49, 388-395.

Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E., and Clark, F. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, 55, 570-582.

Sedlmeier, P. (1999) *Improving Statistical Reasoning: Theoretical Models and Practical Implications*. Mahwah: Lawrence Erlbaum Associates.