# Multipoint Scales: Mean and Median Differences and Observed Significance Levels

**James R. Lewis**

IBM Corporation
Boca Raton, FL

Researchers in human–computer interaction (HCI) often use discrete multipoint scales (such as 5- or 7-point scales) to measure user satisfaction and preference. Many knowledgeable authors state that the median is the appropriate measure of central tendency for such ordinal scales, although others challenge this assertion. This article introduces a new point of view, based on a human factors consideration. When decision makers read a usability report or attend a briefing, they may make decisions based on the magnitude of the difference between the measures of central tendency for key dependent variables. A major criterion that should affect the choice of presenting means or medians is the strength of the relationship between this difference and the observed significance levels of appropriate statistical tests. The results from two series of "real-world" usability studies showed that the mean difference correlated more than the median difference with the observed significance levels (both parametric and nonparametric) for discrete multipoint scale data. Therefore, for these scales in this measurement context, the mean can be a better measure of central tendency than the median. The results also provided evidence that mean differences for 7-point scales correlate more strongly with observed significance levels than those for 5-point scales.

## INTRODUCTION

"There is, of course, nothing strange or scandalous about divisions of opinion among scientists. This is a condition for scientific progress" (Grove, 1989, p. 133).

Researchers in human–computer interaction (HCI) often use discrete multi-

Correspondence and requests for reprints should be sent to James R. Lewis, Human Factors Group, IBM Corp., P.O. Box 1328, Boca Raton, FL 33429-1328.

point scales (such as 5- and 7-point scales) to measure user satisfaction and preference. For example, about one third of the articles in the 1992 *International Journal of Human–Computer Interaction* described measurements that used such scales and reported results using means or statistical procedures that depend on calculating means (Carayon-Sainfort, 1992; Davis & Bostrom, 1992; Henning, Sauter, & Krieg, 1992; Piotrkowski, Cohen, & Coray, 1992; Westlander, & Aberg, 1992; Zapf, Brodbeck, Frese, Peters, & Prumper, 1992). The prevailing belief, based primarily on the principle of invariance (Stevens, 1959), is that the appropriate measure of central tendency for such ordinal measures is the median (Emory, 1976; Klugh, 1970; Townsend & Ashby, 1984). However, others have challenged this assertion often since Stevens first proposed it (Blalock, 1972; Harris, 1985; Lord, 1953; Mueller, Schuessler, & Costner; 1977; Nunnally, 1978), and the controversy has not been resolved (Davison & Sharma, 1988; Davison & Sharma, 1990; Townsend, 1990).

Most practitioners of experimental and applied psychology are familiar with Stevens's (1959) classification of measurement scales (from the principle of invariance) into nominal, ordinal, interval, and ratio, and his rules regarding permissible statistical manipulations:

> The basic principle is this. Having measured a set of items by making numerical assignments in accordance with a set of rules, we are free to change the assignments by whatever group of transformations will preserve the empirical information contained in the scale. These transformations, depending on which group they belong to, will upset some statistical measures and leave others unaffected. In other words, for guidance in setting bounds on the statistical treatment of empirical measurements, we must look to the principle of invariance. The empirical operations that underlie the scale determine what transformations can be made without the sacrifice of information, and the permissible transformations determine, in turn, the appropriate statistical measures, i.e., those that preserve the requisite invariance. (p. 30)

According to Stevens's point of view, multipoint scales are ordinal, and, therefore, the arithmetic mean is not a permissible measure of central tendency for descriptive purposes because adding and dividing numbers from an ordinal scale violates the principle of invariance. Consequently, any statistical procedure that depends on the calculation of the mean is inapplicable to such data.

Psychologists who specialize in measurement and statistics do not universally accept this point of view. Many authors appeal to studies of statistical robustness to justify the application of interval-level statistical methods to ordinal data, although robustness is a poorly defined concept (Bradley, 1978). Nunnally (1978) offered the argument that monotonic transformations of scales (such as 7-point scales) rarely affect inferential decisions. Harris (1985) argued that the level of measurement must affect a researcher's conclusions and generalizations, but not affect permissible arithmetic operations.

When decision makers (who may not understand statistical tests) read a usability report or attend a briefing, they may make decisions based on the magnitude of the difference between the measures of central tendency for key dependent

variables. A major criterion that should affect the choice of presenting means or medians is the strength of the relationship between this difference and the observed significance levels of appropriate statistical tests. In this article, I have presented evidence from two "real-world" usability studies that, for discrete multipoint scales, mean differences correlate more than median differences with observed significance levels (both parametric and nonparametric). This strongly suggests that, from a human factors perspective, HCI researchers should report means rather than medians when they collect data with multipoint scales. (Note that the issue here is the most appropriate measure of central tendency to report, not whether the appropriate test is parametric or nonparametric.)

## THE USABILITY STUDIES

### Office Application System Studies

Lewis, Henry, and Mack (1990) conducted studies to evaluate the usability characteristics of three office application systems. Forty-eight employees of temporary help agencies participated in the studies. The sample size for two of the office systems was 15, and the sample size for the remaining office system was 18. The three office systems included a word processor, a mail application, a calendar application, and a spreadsheet on three different platforms (computer hardware and operating systems) that allowed a certain amount of integration among the applications. We assessed the systems with a set of office benchmark scenarios (Lewis et al., 1990). Two of the system studies had 11 scenarios in common, and the third had 8 scenarios in common with the first two. We used the 3-item After-Scenario Questionnaire (Lewis, 1991a) and the 18-item Post-Study System Usability Questionnaire (Lewis, 1992) to measure participant attitude. All of the questionnaire items were 7-point scales. We used a between-subjects design, assigning each participant to one system. After a 30-min system-exploration period, participants performed the scenarios with their assigned system and completed the After-Scenario Questionnaire (ASQ) after each scenario. After participants completed all scenarios, they rated the system with the Post-Study System Usability Questionnaire (PSSUQ).

### Printer Studies

I conducted scenario-based studies to evaluate the usability of seven printers. Ten different users per printer participated in the studies. The participants worked for temporary employment agencies. I observed participants individually. Each participant performed scenarios in the same order and performed each scenario three times (with minor variations). The seven studies contained four common scenarios (load paper, run the self-test, continue a print job, change the ribbon). After each

scenario, the participants completed a 5-point scale version of the ASQ to indicate their satisfaction (Lewis, 1991b).

## RESULTS

First, I performed all possible pairwise comparisons on the data from the three office application systems for each ASQ item (7-point scale version) and each scenario that was common to a pair of systems. Two systems had 11 common scenarios, and the third system had 8 common scenarios with each of the other two systems. Thus, there were 11 + 8 + 8, or 27, opportunities to compare average ASQ item scores. The total number of comparisons was these 27 opportunities × 3 items per ASQ, for 81 comparisons. Note that three independent means (for those cases in which a scenario was common to all three systems) do not allow for three independent mean comparisons, but they do allow three opportunities to form a relationship among mean differences, median differences, and observed significance levels (OSLs) of applicable parametric and nonparametric tests.

I did the same for each of the items from the ASQ (5-point scale version) in the printer studies. With seven printers, the total number of printer comparisons was $7 \times 6 \div 2$, or 21 printer pairs × 4 common scenarios × 3 trials per scenario × 3 items per ASQ, for a total of 756 opportunities to examine the relationship among mean differences, median differences, and OSLs.

I calculated the means, the mean difference, the OSL from an independent groups *t* test, the medians, the median difference, and the OSL from a Mann-Whitney *U* test for each comparison. (See Lewis, 1989, for a comprehensive treatment and listing of the results for the office application systems.) The independent groups *t* test evaluates the significance of mean differences and the Mann-Whitney *U*-test is a nonparametric test that researchers can use to assess the median difference (Bradley, 1976).

For both sets of data, I obtained 95% confidence intervals (CIs) for the product-moment correlations (Steele & Torrie, 1960) among the mean differences, median differences, *t* test OSLs, and *U* test OSLs, shown in Table 1. The correlations for the office applications studies have 79 degrees of freedom (81 pairwise comparisons among three systems −2), and those for the printer studies have 754 degrees of freedom (756 pairwise comparisons among seven printers −2). The pattern of correlations was consistent across the two data sets. Although all correlations were significant ($p < .05$), the 95% CIs showed that the mean difference was significantly ($p < .05$) more related than the median difference to both the *t* test OSL and the *U* test OSL.

The confidence intervals for the correlations between the mean difference and the *t* test OSL for the printers study and the office applications study did not overlap. Because this result could be an artifact of the difference in sample sizes between the studies, I deleted portions of each study to equate the sample sizes. (This

**Table 1. Mean, median, and OSL correlation analyses (all possible comparisons)**

| | Office Applications Study (7-Point Scales, $N = 81$) | | |
|---|---|---|---|
| | *t* test OSL (95% CI) | Median Difference (95% CI) | *U* test OSL (95% CI) |
| Mean Difference: | −0.92 (−0.88, −0.94) | 0.54 ( 0.39, 0.65) | −0.89 (−0.86, −0.92) |
| *t* test OSL: | | −0.50 (−0.33, −0.59) | 0.96 ( 0.92, 0.97) |
| Median Difference: | | | −0.53 (−0.38, −0.64) |
| | **Printers Study (5-Point Scales, $N = 756$)** | | |
| | *t* test OSL (95% CI) | Median Difference (95% CI) | *U* test OSL (95% CI) |
| Mean Difference: | −0.80 (−0.75, −0.82) | 0.54 ( 0.49, 0.61) | −0.80 (−0.75, −0.82) |
| *t* test OSL: | | −0.27 (−0.16, −0.34) | 0.90 ( 0.88, 0.91) |
| Median Difference: | | | −0.44 (−0.37, −0.52) |

*Note.* OSL = observed significance level; CI = confidence interval.

also makes the use of OSLs as an indicator of effect size more defensible.) Participants in the office applications study only attempted each scenario once, so I deleted the second and third trials from the printers study. Because the printers study had only four scenarios in common across all printers, I randomly deleted four of the eight common office application scenarios and also deleted the three scenarios that were common to only two systems. The printers study only had 10 participants per printer, so I randomly selected 10 participants to retain for each of the systems in the office applications study. I then analyzed these data to obtain the 95% CIs for the two studies given equal sample sizes, shown in Table 2 (p. 388). The pattern of results is similar to that of Table 1. Although the intervals were closer, the CIs for the correlations between the mean difference and the *t* test OSL for the printers study and the office applications study still did not overlap.

Finally, I examined a subset of the data composed of independent multiple comparisons to ensure that the previous patterns of results were not distorted by any violations of assumptions of independence among comparisons. I restricted the office applications comparisons to the two systems that had 11 common scenarios. The independent data base included the 33 ASQ averages (11 scenarios × 3 items per ASQ) and the 18 items from the PSSUQ for a total of 51 independent opportunities to assess the relationship among mean differences, median differences, and OSLs. For the printers study, I randomly paired six of the seven printers (4 with 5, 2 with 6, 1 with 3) to obtain a data base that contained 108 independent comparisons (3 sets of printer pairs × 4 common scenarios × 3 trials × 3 items per ASQ). As shown in Table 3 (p. 388), the pattern of correlations (and 95% CIs for correlations) is virtually the same as those shown in Tables 1 and 2, which demonstrates that these patterns are not an artifact caused by violations of assumptions of independence.

**Table 2. Mean, median, and OSL correlation analyses (equal number of comparisons)**

| | Office Applications Study (7-Point Scale, $N = 36$) | | |
|---|---|---|---|
| | *t* test OSL (95% CI) | Median Difference (95% CI) | *U* test OSL (95% CI) |
| Mean Difference: | −0.94 (−0.92, −0.95) | 0.64 ( 0.46, 0.77) | −0.84 (−0.74, −0.90) |
| *t* test OSL: | | −0.54 (−0.33, −0.70) | 0.96 ( 0.92, 0.97) |
| Median Difference: | | | −0.48 (−0.23, −0.66) |
| | **Printers Study (5-Point Scales, $N = 36$)** | | |
| | *t* test OSL (95% CI) | Median Difference (95% CI) | *U* test OSL (95% CI) |
| Mean Difference: | −0.85 (−0.75, −0.91) | 0.34 ( 0.09, 0.56) | −0.83 (−0.73, −0.89) |
| *t* test OSL: | | −0.15 (−0.40, 0.13) | 0.86 ( 0.76, 0.92) |
| Median Difference: | | | −0.32 (−0.07, −0.54) |

*Note.* OSL = observed significance level; CI = confidence interval.

**Table 3. Mean, median, and OSL correlation analyses (independent comparisons)**

| | Office Applications Study (7-Point Scales, $N = 51$) | | |
|---|---|---|---|
| | *t* test OSL (95% CI) | Median Difference (95% CI) | *U* test OSL (95% CI) |
| Mean Difference: | −0.92 (−0.89, −0.95) | 0.55 ( 0.35, 0.61) | −0.90 (−0.84, −0.93) |
| *t* test OSL: | | −0.45 (−0.20, −0.64) | 0.97 ( 0.96, 0.98) |
| Median Difference: | | | −0.51 (−0.27, −0.67) |
| | **Printers Study (5-Point Scales, $N = 108$)** | | |
| | *t* test OSL (95% CI) | Median Difference (95% CI) | *U* test OSL (95% CI) |
| Mean Difference: | −0.78 (−0.68, −0.84) | 0.42 ( 0.23, 0.57) | −0.77 (−0.67, −0.83) |
| *t* test OSL: | | −0.25 (−0.05, −0.42) | 0.86 ( 0.80, 0.90) |
| Median Difference: | | | −0.40 (−0.21, −0.55) |

*Note.* OSL = observed significance level; CI = confidence interval.

## DISCUSSION

When researchers present the results of a usability study, they must consider the impact that the reported averages and average differences will have on the decision makers. Therefore, the differences between reported averages should correlate strongly with the OSL of the statistical tests used to assess the average difference.

The correlations between the OSLs for the *t* tests and the Mann-Whitney *U*

tests showed substantial agreement for both sets of data. The correlations between the mean difference and the OSLs for both the *t* tests and the Mann-Whitney *U* tests were also substantial, showing that as the mean difference increased the OSLs for both the parametric and nonparametric tests decreased (became more significant). The correlations between the median difference and the statistical tests were relatively low. These results showed that, for discrete multipoint scales in this measurement context, the mean difference correlated more than the median difference with the OSLs (both parametric and nonparametric).

Why is the median difference such a poor indicator of statistical significance? In part, these results may relate to the fact that sample means drawn from a continuous distribution are less variable than sample medians (Blalock, 1972). For these types of scales, however, there are two additional factors to consider. First, note that the median of a multipoint scale must have a discrete distribution regardless of the sample size, but the distribution of the mean will become more continuous as the sample size increases. It follows that the distribution of the median difference will also be discrete regardless of sample size, but the distribution of the mean difference will become more continuous as the sample size increases. Because the mean difference can acquire a larger number of values, it can reflect significant differences between the samples more reliably than the median difference. When scales are open-ended (have at least one endpoint at infinity) and sample sizes are small, extreme values affect means but do not affect medians. However, multipoint scales are not open-ended, so the median does not have an advantage over the mean in this respect.

The paired comparisons used sample sizes that ranged from 25 (with missing data) to 33 (15 + 18 participants) for the office applications studies and from 13 (with missing data) to 20 (10 + 10 participants) for the printers studies. It is likely that these results generalize to all sample sizes. As sample sizes increase, the mean becomes more continuously distributed and the median remains discrete. Also, according to the central limit theorem, as the sample size increases the sampling distribution of the mean approaches a normal distribution (Bradley, 1976), although the characteristics of the sampled population dictate the speed of this approach (Bradley, 1973). Thus, the central limit theorem suggests that the mean difference will become a better indicator of population difference as the sample size increases. It is possible that as the sample size decreases, there might be a point at which the median difference would become a more reliable indicator of statistical difference than the mean difference, but this is not likely. It is more likely that mean and median differences both become less reliable indicators of statistical significance as the sample size decreases, with the mean difference reliability decreasing at a relatively faster rate until the mean and median become the same measure of central tendency when the sample size is two.

Because these scales are discrete with a limited response range, the results should generalize to discrete scales with fewer than 5 points. When the scale has more than 7 response points, however, it is hard to determine the limits of generalization because the scale may eventually reach a point at which a single extreme score could unduly affect the mean. That probably will not happen with 8 or 9 points, but may happen with 100 or 1,000 points. However, few researchers use

scales with this many points because psychometric research has shown that increasing the number of scale steps increases scale reliability rapidly from 2 to 7 steps, but with little gain in reliability after 11 steps (Nunnally, 1978). The results of the current study are consistent with this psychometric research because the means of the 7-point scales correlated more than the means of the 5-point scales with the OSLs of the *t* tests.

It is difficult to predict what would happen with a scale that allowed participants to make a continuous response. If the scale was continuous, the median difference would no longer follow a discrete distribution and could be as reliable an indicator of significance as the mean difference. On the other hand, Blalock (1972) stated that even for continuous interval data, sample medians are usually more variable than sample means.

In summary, the following three criteria, in order of priority, are pertinent to the choice of the best central value (Mueller et al., 1977). First, what is the purpose of the average? Second, what is the pattern of the distribution of the data? Third, what are the technical (primarily arithmetic) considerations that affect the choice of average?

In the context of presenting results to decision makers, the primary purpose of the averages is to enhance decision making. Therefore, it is reasonable to choose an average for which the differences between averages being compared corresponds to the likelihood that these differences did or did not occur by chance (observed significance level). Distributional patterns can be flat or peaked, symmetric or skewed. A consideration that generally affects the choice of an average is the skewness of the distribution. The more skewed a distribution, the better the median represents its central tendency. However, a technical consideration that reduces the importance of skewness is whether a distribution has open-ended intervals. A distribution with close-ended intervals (such as a 5- or 7-point scale) will not necessarily benefit from reporting the median in preference to the mean. An additional technical consideration is that the median (and median differences) is discrete regardless of sample size, but the mean (and mean differences) becomes more continuous as the sample size increases. It is also important to note the similarity of results between two studies that differed in measuring instruments (5-point vs. 7-point scales) and contexts of use (office application scenarios vs. printers scenarios), providing evidence for generalization. Considering all the evidence, it appears that, for this measurement context and purpose of presentation, it is better to report the mean than the median.

## CONCLUSIONS

For multipoint scales, the mean difference correlated more than the median difference with the OSLs (both parametric and nonparametric) for discrete multipoint scale data. The mean difference of a 7-point scale correlated more than the mean difference of a 5-point scale with the OSLs of *t* tests.

As an aid to the decision makers who read usability reports, researchers should use the mean rather than the median to report the central tendency of the

data they collect with multipoint scales. Also, researchers should use 7-point scales rather than 5-point scales for these measurements.

## REFERENCES


Blalock, H. M. (1972). *Social statistics.* New York: McGraw-Hill.

Bradley, J. V. (1973). The central limit effect for a variety of populations and the influence of population moments. *Journal of Quality Technology, 5,* 171–177.

Bradley, J. V. (1976). *Probability, decision, statistics.* Englewood Cliffs, NJ: Prentice-Hall.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144–152.

Carayon-Sainfort, P. (1992). The use of computers in offices: Impact on task characteristics and worker stress. *International Journal of Human–Computer Interaction, 4,* 245–261.

Davis, S., & Bostrom, R. (1992). An experimental investigation of the roles of the computer interface and individual characteristics in the learning of computer systems. *International Journal of Human–Computer Interaction, 4,* 143–172.

Davison, M. L., & Sharma, A. R. (1988). Parametric statistics and levels of measurement. *Psychological Bulletin, 104,* 137–144.

Davison, M. L., & Sharma, A. R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regression. *Psychological Bulletin, 107,* 394–400.

Emory, C. W. (1976). *Business research methods.* Homewood, IL: Richard D. Irwin.

Grove, J. W. (1989). *In defence of science: Science, technology, and politics in modern society.* Toronto: University of Toronto Press.

Harris, R. J. (1985). *A primer of multivariate statistics.* Orlando, FL: Academic.

Henning, R. A., Sauter, S. L., & Krieg, E. F. (1992). Work rhythm and physiological rhythms in repetitive computer work: Effects of synchronization on well-being. *International Journal of Human–Computer Interaction, 4,* 233–243.

Klugh, H. E. (1970). *Statistics: The essentials for research.* New York: Wiley.

Lewis, J. R. (1989). *The relative reliabilities of mean and median differences as indicators of statistically significant differences for 7-point scales* (Report No. 54.532). Boca Raton, FL: IBM Corporation.

Lewis, J. R. (1991a). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ. *SIGCHI Bulletin, 23,* 78–81.

Lewis, J. R. (1991b). An after-scenario questionnaire for usability studies: Psychometric evaluation over three trials. *SIGCHI Bulletin, 23,* 79.

Lewis, J. R. (1992). Psychometric evaluation of a post-study system usability questionnaire: The PSSUQ. *Proceedings of the Human Factors Society 36th Annual Meeting.* Santa Monica, CA: Human Factors Society.

Lewis, J. R., Henry, S. C., & Mack, R. L. (1990). Integrated office software benchmarks: A case study. *INTERACT '90.* London: North-Holland.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750–751.

Mueller, J. H., Schuessler, K. F., & Costner, H. L. (1977). *Statistical reasoning in sociology.* Boston: Houghton Mifflin.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Piotrkowski, C. S., Cohen, B. G. F., & Coray, K. E. (1992). Working conditions and well-being among women office workers. *International Journal of Human–Computer Interaction, 4,* 263–281.

Steele, R. G. D., & Torrie, J. H. (1960). *Principles and procedures of statistics.* New York: McGraw-Hill.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theory.* New York: Wiley.

Townsend, J. T. (1990). Truth and consequences of ordinal differences in statistical distributions: Toward a theory of hierarchical inference. *Psychological Bulletin, 108,* 551–567.

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin, 96,* 294–401.

Westlander, G., & Åberg, E. (1992). Variety in VDT work: An issue for assessment in work environment research. *International Journal of Human–Computer Interaction, 4,* 283–301.

Zapf, D., Brodbeck, F. C., Frese, M., Peters, H., & Prumper, J. (1992). Errors in working with office computers: A first validation of a taxonomy for observed errors in a field setting. *International Journal of Human–Computer Interaction, 4,* 311–339.