This paper won the David Hardin Award from the American Marketing Association for the Best Paper to appear in Marketing Research Magazine in 2004.

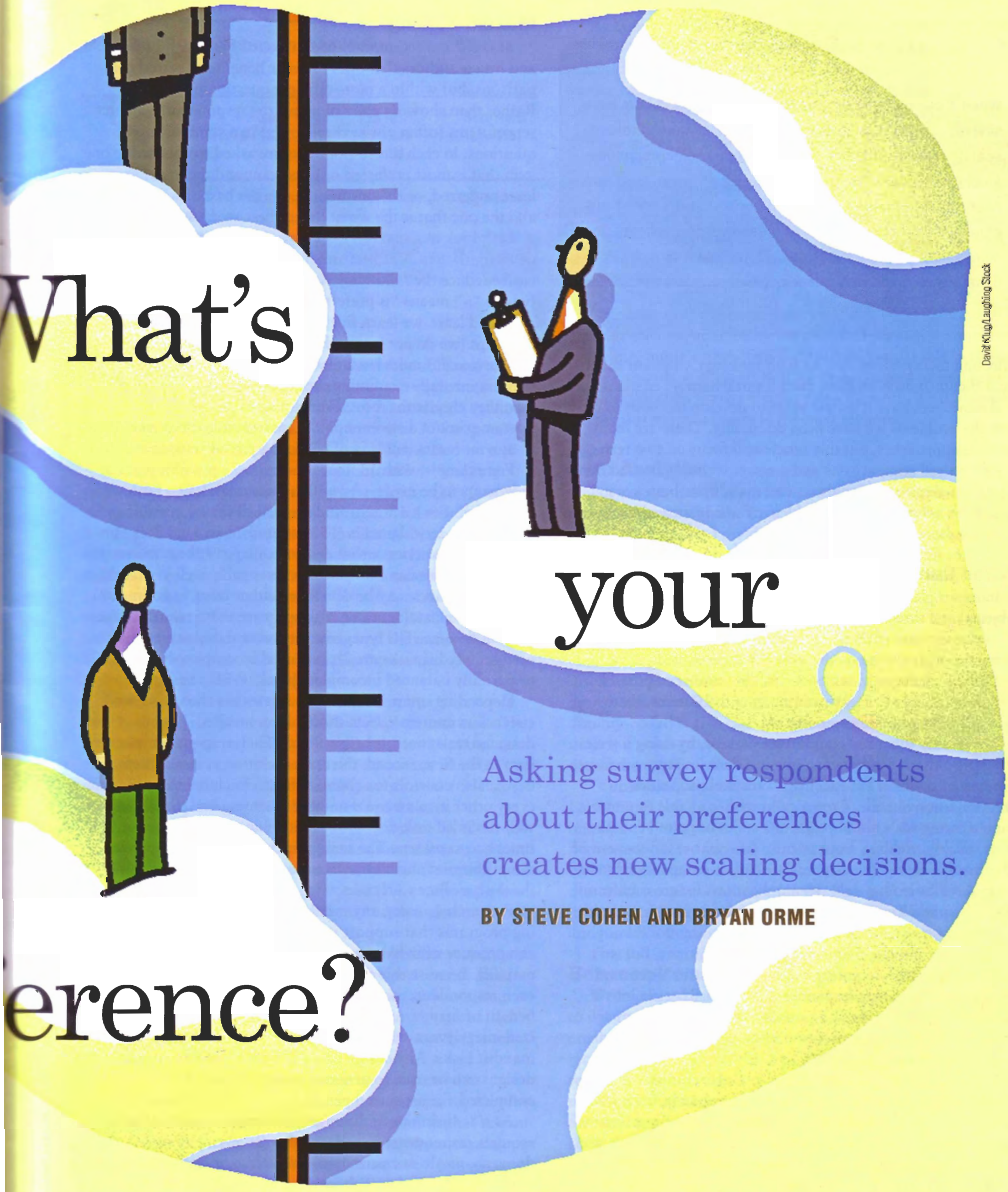
Common tasks for market researchers include measuring people's preferences for brands, flavors, or colors and measuring the importance of product features and benefits. For these kinds of problems, the monadic rating remains a very popular approach. A typical study has respondents rate a series of items on a 5- or 10-point desirability or importance scale, often with a dozen—or even several dozen—items being investigated. From an analysis of these data, the researcher makes recommendations about marketing and product strategy. This method has been time-tested and forms the basis of many research studies.

However, we also know that respondents are notorious for rating items very rapidly, using simplification heuristics to speed through the task. How can we blame them, when we present so many items for their consideration? So many studies show that respondents use only a limited range of the scale points, resulting in many ties across items. Thus, recommendations are made often on trivial statistical differences. Moreover, the literature is replete with examples of people evidencing response style biases. Some respondents use just the top few boxes of a rating scale, some refuse to register a top score for any item, while others conscientiously spread their ratings across the entire range. Response style bias poses problems for statistical tests of significance and multivariate modeling. Separating the "signal" from the noise is difficult.

Too often we ignore these issues and present the raw ratings. While standardization of ratings (forcing the mean rating within each respondent to zero and the standard deviation to unity) has often been suggested as an appropriate remedy, this transformation removes the level differences between respondents and is often difficult for managers to understand. Furthermore, when a respondent uses just a few scale points, the within-respondent standard deviation is very small, making the new standardized estimate very large. These deficiencies lead to the rare use of standardization in commercial practice.

To ameliorate the situation of low discrimination across items, some researchers use rankings. In a ranking task, respondents order the items from best to worst (with no ties). Hence, respondents must discriminate among items. However, respondents often find it difficult to rank more than about seven items. So test-retest reliability of long ranked lists tends to be low.

Researchers have experimented with many techniques to achieve the benefits of metric scaling while also encouraging respondents to discriminate among the items. Another common approach is the "constant sum" or chip allocation scale. To use a constant sum scale, respondents allocate a certain number of points or chips across an array of items. Some researchers declare that the constant sum scale is a more discriminating task, but it also has its limitations. Respondents find it difficult to make the allocations sum to the required number, and this hurdle may get in the way of respondents accurately



What's your erence?

Asking survey respondents about their preferences creates new scaling decisions.

BY STEVE COHEN AND BRYAN ORME

David Klug/Laughing Stock

Executive Summary

When it comes to scaling multiple items, researchers have several options. This article compares maximum difference scaling against monadic ratings and paired comparisons. Among other benefits, it is flexible enough to be used with paper questionnaires and computerized interviewing programs. It also offers improved measures of discrimination across items over rating scales. The article discusses additional benefits and weaknesses of all three methods.

reflecting their preferences. As with rankings, constant sums are difficult to do with more than a small number of items.

Researchers have tried for years to develop methods to deal with the problems we have been discussing. There are many possible approaches, but this article will focus on two techniques: paired comparisons and a newer technique called maximum difference scaling (also called maxdiff or best-worst scaling) as solutions to the problem of sturdy measurement of preferences.

Paired Comparisons

In its simplest form, respondents are shown just two items (objects) at a time and asked to choose between them. Consider the case of three items: A, B, and C. With three items, there are three possible comparisons: A vs. B; A vs. C; and B vs. C. Generally, with t items, the number of possible comparisons is $\frac{1}{2}\,t\,(t\text{-}1)$. For example, with 10 items, there are $\frac{1}{2}\,(10)(9) = 45$ possible comparisons.

The beauty of paired comparisons is that, by using a series of simple either/or judgments, one can derive a strong measure of preference which has excellent statistical properties on a common interval scale. Even a child who is unable to understand a rating scale could perform a series of paired comparisons reliably, yielding a statistically informative assessment of all the items. Being able to translate simple comparisons to an interval scale also has valuable implications for cross-cultural research, where controlling for response style bias is desirable.

Paired comparison questions indeed require more thought and time to complete than simple monadic ratings. But isn't that a good thing—requiring thinking to answer a question? With monadic ratings, respondents more easily settle into a less motivated, patterned-response strategy. Paired comparisons thus discourage inattention and seem to elicit more discerning responses.

With many items, the number of possible comparisons between pairs can become very large. Fortunately, it's not necessary to ask respondents to make all possible comparisons to obtain reasonably stable interval-scale estimates for each item. Incomplete (fractional) designs that require showing just a carefully chosen subset of the universe of comparisons are more than adequate in practice.

Maxdiff Scaling

Maxdiff is a recent method developed by Jordan Louviere and his co-authors that offers similar benefits as paired comparisons, but within a more efficient questioning method. Rather than showing pairs of items, the respondent evaluates sets of items (often of size three to six) in a series of choice questions. In each set, respondents are asked to choose the one item that is most preferred or important and the one that is least preferred, or the one item that is the best of something and the one that is the worst (hence best-worst scaling).

Let's look at a maxdiff question with four items. If respondents mark item A as best and item D as worst, then we can easily deduce the following: A>B, A>C, A>D, B>D, C>D (where ">" means "is preferred to"). Notice that by asking the most and least, we learn five of the six possible paired relationships (we do not learn whether B is preferred to C).

The maxdiff model assumes that respondents behave as if they are mentally examining every possible pair in each set, and then they choose the most distinct pair as the best-worst, most-least, maximum difference pair. Thus, one may think of maxdiff as a more efficient way of collecting paired comparison data.

For example, with 13 items, there are 78 possible pairs or judgments to be made, and with a fractional design far fewer are needed. For a 13-item study, a well-designed maxdiff task will employ 26 judgments (13 bests and 13 worsts). It is our commercial experience that designs with fewer best/worst tasks than items can also perform reasonably well in practice.

Maxdiff tasks may be developed either using experimental designs from catalogues or with software that generates experimental designs. The types of experimental designs most often used in maxdiff tasks are 2^k, balanced incomplete block (BIB), or partially balanced incomplete block (P-BIB) designs.

Depending upon which design strategy is chosen, order effects and context effects should be controlled. A well-designed task controls order effects: Each respondent sees each item in the first, second, third, etc. position across subsets. The design also controls for context effects: Each item is seen with every other item an equal number of times. And, finally, a well-designed task displays each item an equal number of times across all sets. The same principles that govern traditional conjoint and choice-based conjoint tasks also apply to the design of maxdiff tasks.

Researchers using any number of computerized interviewing programs that support item-based randomization logic can produce near-balanced plans for paired comparison or maxdiff. Because this approach leads to unique interviews for each respondent, these "randomized" plans have the added benefit of further reducing order effects. However, computerized interviews are not necessary for the construction of maxdiff tasks. Any design catalogue that lists BIB or PBIB designs can be used to generate maxdiff tasks that can be completed via paper and pencil.

As a "cousin" to traditional conjoint analysis, maxdiff requires respondents to make trade-offs among the items. By doing so, we do not permit anyone to like or dislike all items. By definition, the act of choosing a most and a least elicits the relative preferences out of the respondent.

Exhibit 1 Time to complete exercise

	Monadic ($n=137$)	Paired Comparison ($n=121$)	Best/Worst ($n=116$)
Mean time to complete exercise	97 seconds	320 seconds	298 seconds
Seconds per mouse click	4.9 sec./click	10.7 sec./click	9.9 sec./click

Comparing Methods

Together with Michael Patterson, formerly of Hewlett-Packard, we conducted a methodological experiment, focusing on measuring the importance of a list of 20 attributes related to the purchase of file servers. We interviewed 374 respondents from a list provided by HP, using a Web-based survey.

The sample was randomly divided into thirds, and respondents were asked to indicate their preference for the 20 items using either monadic importance ratings on a 1-9 scale, where 1 indicated "not important" and 9 indicated "extremely important" (n=137), paired comparisons (n=121), or maximum difference scaling (n=116). (Our monadic scale may not have been the "best" application of a ratings scale, but we think it is quite reflective of general practice.) Additional holdout tasks were asked at the beginning and the end of the interview, wherein respondents ranked the importance of three items in each of four different sets.

Because we were interviewing using computers, we could time how long it took respondents to complete each task. The monadic ratings required 20 mouse clicks. For the paired comparison experiment, we showed each respondent 30 pairs, which required 30 clicks. For maxdiff, we showed 15 sets of four items, requiring two clicks per set or 30 clicks.

On average, the paired comparisons and maxdiff task took about three times as long to complete as the ratings as seen in Exhibit 1. On the basis of "seconds per click," the ratings task took about half as long as the other two tasks, indicating less to read and, perhaps, less involvement and thought for the easier monadic task.

After the interview was complete, we asked the respondents to tell us their perceptions of the task they performed. As can be seen in Exhibit 2 on page 37, on a 7-point scale of disagree-agree, all tasks were evaluated at about the midpoint of each scale, with ratings being rated slightly better than the paired comparison or maxdiff tasks.

Although the monadic rating exercise was viewed as slightly more enjoyable, less confusing, and easier, the absolute differences are quite small. The important point is that respondents did not perceive any of the tasks to be very confusing or difficult. We think these results demonstrate that paired comparisons and maxdiff scaling are quite doable in surveys and are a good alternative to traditional monadic ratings.

Although we do not present the data here, the rank order of the importance of the items yielded similar results across the methods. For example, see Exhibit 3 on page 37. Both axes have a range of eight points. The mean ratings for each of the 20 items are shown in the scatterplot. Three things are of note:

1. The rank correlation between the sets of items is high (more than 0.80). This indicates, as just stated, that the two methods yielded comparable overall results.

2. The range of the mean ratings is only about two points. With one exception, all ratings have a mean score between six and eight. With such small differences, we might not expect tests of differences to be statistically significant.

3. The range of the maxdiff scores is much larger: about five points. This is affected by the choice of dummy coding and scaling, but if within-item variation is comparable between the methods, it suggests that we might expect to find greater discrimination between items using maxdiff. We'll investigate this more formally in the next sections.

Data Analysis

For the monadic ratings, we can use the ratings themselves or we can normalize the data within respondent (by subtracting off the mean and dividing by the standard deviation). We generally used the raw ratings for our data analyses and resorted to the standardized ratings for one of our tests.

For paired comparisons, if a graded rating scale is used (degree of preference for left over right), ordinary least squares regression can be used to estimate individual level scores. If choices are indicated (prefer left over right), a simple binary discrete choice model may be used to analyze the data. The design matrix indicating which items were compared was dummy-coded in this case. As an alternative to the binary logit model, it's also possible to use Thurstone case V scaling to derive the required scale values.

For maxdiff, discrete choice models are used with a twist. Each choice task is coded twice: once for the chosen item and once for the rejected item. The design matrix shows which items were in the task and is flipped (all values multiplied by -1) for the rejected choice sets. The analyses presented in this article employs estimated individual-level models using hierarchical Bayes choice models for the paired comparison and maxdiff data.

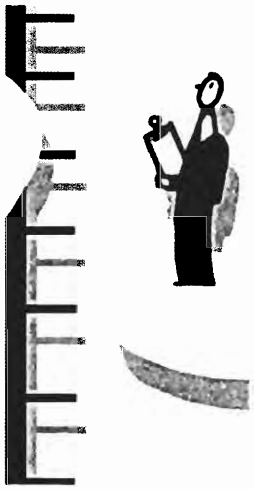
Between-Item Discrimination

We hypothesized that the methods that forced respondents to discriminate among items should result in measures of preference that reflect greater discrimination than standard monadic scales. To test this hypothesis, we performed a repeated measures ANOVA within each of the three methods, requesting the Student-Newman-Keuls test of post hoc differences. This test compares all possible pairs of items (20×19/2=190 pairs), adjusting the results to reflect the fact that the researcher has performed multiple tests of significance. All ANOVA results indicated that there were differences in mean ratings across items, although the number of significant differences differed by method. For the ratings data, only 44% of the post hoc t-tests were found to be statistically significant. For the pairs data, 83% of the tests were significant, and for the maxdiff data, 76% of the tests were significant. We conclude that the rating scale discriminated least among the items when comparing each one to the other, and that the paired comparison task performed slightly better than maxdiff.

With the server study, we included some holdout tasks. At the very beginning and at the very end of the survey, we showed each respondent four sets of three items. We then asked each respondent to rank-order the three items in each of the four sets. We computed the test-retest reliability of the two ranking tasks. Then, using the monadic ratings, paired comparisons, and maxdiff importance scores, we predicted the preference for the three pairs implied in the holdout ranking tasks.

Relative to what was obtained with test-retest reliability, the predictive hit rates were 85%, 88%, and 97% for monadic ratings, paired comparisons, and maxdiff, respectively. While the performance of paired comparisons and ratings is commendable, the maxdiff performance is excellent, performing at about the same level as test-retest reliability.

Even a child who is unable to understand a rating scale could perform a series of paired comparisons reliably, yielding a statistically informative assessment of all the items.



Although we do not present the data here, we found that maxdiff could still predict holdout pairs (relative to test-retest reliability) as well as paired comparisons even after discarding the second half of the maxdiff tasks for each respondent. This confirms our earlier observation that maxdiff questions are more efficient than paired comparisons.

Between-Group Discrimination

Any good researcher will want to know how well these measures did in discriminating across respondent groups. In the survey, we included a number of background questions, including product usage, attitudes, and behaviors in addition to the importance measurement section. Using these questions, we constructed 19 two-category variables (e.g., high/low product use; bigger/smaller companies). With the preferences of the 20 items in hand, we performed 380 separate tests of differences across groups (380 = 19 background variables × 20 importance ratings). At the 95% confidence level, we would expect to find 19 significant differences by chance alone.

We found 30 significant differences between respondent groups for the monadic ratings. To see if within-respondent standardization made any difference, we re-ran these tests against within-respondent standardized scores and found only 22 significant differences (suggesting that some of the differences seen may have been due to response style bias). Running the same tests with the other two methods, we obtained 40 significant differences with the paired comparisons data, and 37 with maxdiff. Once again, we conclude the rating scales are less discriminating than the other two methods, and again paired comparisons performed just slightly better than maxdiff.

Dealing With the Relative Scale

Depending on your viewpoint, one weakness of paired comparisons or maxdiff is that the scores reflect relative, rather than absolute, preferences. The derived scores are based on the relative comparisons among the items included in the study and will change if the content or number of items being compared changes.

Since the paired comparison and the maxdiff method utilize choice responses, we can easily transform the derived scores (assuming the parameters were fit using a logit model specification) to a probability scale. Assuming zero-centered raw scores, one can use the transform:

$$P_i = 100\,[e^u/(1+e^u)]$$

Here, P_i indicates probability of choosing item i, and the u on the right-hand side of the equation is the raw score for the i item.

This transform places the scores on a 0 to 100 scale. The interpretation of a "15" is that this item would be chosen (or not rejected) 15% of the time on average when compared to the other items.

In our experience, not having an absolute score for each item is not a major impediment. Judicious selection of the set of items in the study ensures that the results may be compared across items and used for decision making.

Benefits for Market Segmentation

Although it is beyond the scope of this article, we contend that the use of maxdiff measurement can particularly enhance market segmentation studies. As has been seen, maxdiff enjoys improved measures of discrimination across items over rating scales. The ability to distinguish across items should also permit us to uncover differences across respondents. This ability is a significant improvement over common practice, which typically uses rating scales as the basis variables for segmentation. If the basis variables do not discriminate well, how can we expect segments based on these data to be something other than figments of our statistical imagination? Since segmentation is a search for differences and discrimination, why not use methods that encourage such results?

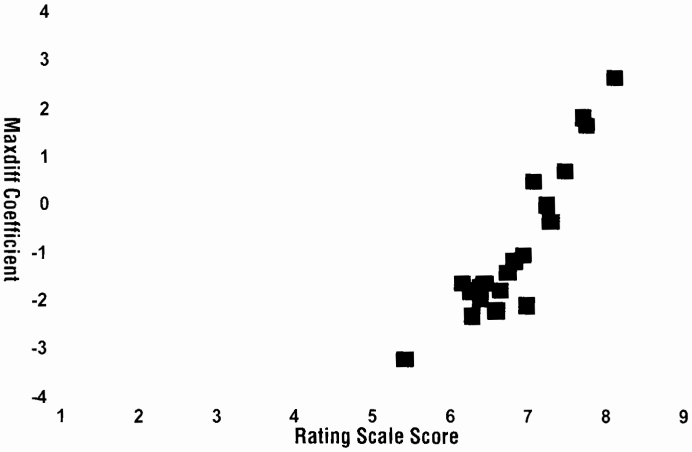
Exhibit 2 Qualitative evaluation of each task

Using a scale where 1 means "strongly disagree" and 7 means "strongly agree," how much do you agree or disagree that the previous section...

	Monadic (*n* = 137)	Paired Comparison (*n* = 121)	Best/Worst (*n* = 116)
...was enjoyable	4.3 (b, c)	4.0 (a)	3.8 (a)
...was confusing	2.4 (b, c)	2.9 (a)	3.2 (a)
...was easy	5.6 (b, c)	5.2 (a)	5.1 (a)
...made me feel like clicking answers just to get done	3.2	3.1 (c)	3.6 (b)
...allowed me to express my opinions	4.9 (c)	4.6	4.3 (a)

a" means significantly different from column a, p<0.05, etc.)

Exhibit 3 Mean results from ratings and maxdiff tasks



Because it does not employ a rating scale, maxdiff also has the advantage that it can be employed in cross-cultural studies with ease. There's no need to worry about the response biases that are so often found in multi-country segmentation studies. In a review of issues and the practice of international segmentation research, Jan-Benedict E.M. Steenkamp and Frenkel ter Hofstede speculate that a major reason that multi-country segmentation studies tend to show a bias toward country-specific results, rather than toward regional or cross-country results, is because of the differential use of rating scales across countries.

Maxdiff tasks can be designed for use both with computerized interviewing programs and with the use of a paper questionnaire, making it applicable across a wide range of research contexts. Because maxdiff can be used with relatively small fixed designs, it has an advantage over paired comparisons when a computerized interview is not possible. Finally, because maxdiff collects choice responses, the data collected are very amenable to the use of modern segmentation methods like latent class choice models.

We collected data from more than 300 IT managers and compared the results of monadic ratings, paired comparisons, and maxdiff scaling with respect to three measures of performance: validity (hit rates), discrimination among items, and discrimination across respondent groups. On all three measures, monadic ratings were found to be inferior to the other two methods. Maxdiff measurement achieved the highest predictive validity. Paired comparisons and maxdiff were much better than monadic ratings at finding significant differences between subgroups and across items. Importantly, maxdiff is much more efficient than paired comparisons, providing more information per level of respondent effort.

Therefore, we suggest that practitioners adopt maximum difference scaling as another method in their toolbox for developing a unidimensional scale of importance or object preference. The maxdiff task is easy for a respondent to do and is scale-free, so it can easily be used to compare results across diverse individuals. Furthermore, it's easy to implement on paper or with a computer, relatively easy to analyze with standard software, and easy to explain to respondents and managers alike.

Additional Reading

Baumgartner, Hans and Jan-Benedict E.M. Steenkamp (2001), "Response Styles in Marketing Research: A Cross-National Investigation," *Journal of Marketing Research*, 38 (May).

David, H. A. (1969), *The Method of Paired Comparisons*, London: Charles Griffin & Co.

DeSarbo, Wayne S., Venkatram Ramaswamy, and Steven H. Cohen (1995), "Market Segmentation With Choice-Based Conjoint Analysis," *Marketing Letters*, 6 (2), 137-147.

Finn, Adam and Jordan J. Louviere (1992), "Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety," *Journal of Public Policy and Marketing*, 11 (1), 19-25.

Louviere, Jordan J. (1992), "Maximum Difference Conjoint: Theory, Methods and Cross-Task Comparisons With Ratings-Based and Yes/No Full Profile Conjoint." Unpublished Paper, Department of Marketing, Eccles School of Business, University of Utah, Salt Lake City.

Steenkamp, Jan-Benedict E.M. and Frenkel Ter Hofstede (2002), "International Market Segmentation: Issues and Outlook," *International Journal of Research in Marketing*, 19, 185-213. ●

Steve Cohen is principal of SHC & Associates and may be reached at steve@shcstrat.com. **Bryan Orme** is vice president of Sawtooth Software Inc. and may be reached at bryan@sawtoothsoftware.com.