

{ Testing Alternatives to Importance Ratings }

By Keith Chrzan, Division Vice President, Marketing Sciences Group, Maritz Research and Natalya Golovashkina, Ph.D., Senior Research Analyst, Maritz Research

Introduction

Many marketing researchers have come to realize rating scales do a poor job of measuring stated importance.

- In the absence of constraints that prevent them from rating all attributes as extremely important, respondents often do just that, making it hard to tell more from less important attributes
- Respondents tend to use rating scales differently: some tend to rate high and some tend to rate low, for example; some use only a narrow band of the scale and some use the whole width of the scale. This simply invalidates the needs-based segmentations that use importance ratings, because real differences between people will be partially confounded by the scale artifact (which academics call “scale use heterogeneity”)
- Moreover, respondents in different countries tend to use rating scales systematically differently, so to the scale artifact that occurs on an individual respondent level you can add an additional artifact that occurs at the cultural level, making the idea of a multinational needs-based segmentation study based on importance ratings a complete disaster
- Finally, and most devastatingly of all, importance ratings have been found to lack predictive validity. When included in a predictive model called “the multi-attribute attitude model,” importance ratings actually make the model predict worse when they are included than when they are removed¹

In a 2006 survey of over 400 applied researchers, 53% of them reported using rating scales to measure attribute importance. Why the disconnect? Perhaps researchers do not

realize how bad rating scales are at measuring importance. Then again, it could be that alternatives to importance rating scales have not been tested and found to be viable. To address both explanations for this disconnect, Maritz Research sponsored a primary research study to test alternative ways of measuring stated attribute importance. The *International Journal of Market Research*² published this study late last year and a summary appears below.

Objectives for Stated Importance Measures

So, what would we expect of a good stated importance measure?

First, if it does nothing else, an importance measure should be able to tell us which attributes are more important than others. In other words, it should have *discriminating power*.

Questionnaire “real estate” is expensive, so, all else being equal, we will prefer an importance measurement that takes less space or time over one that takes more.

Finally, and most importantly, the measure should have predictive power. A product or service that has more of the important attributes should be preferred to one that possesses the important attributes to a lesser degree.

Research Design

To compare several alternatives to standard importance ratings, Maritz fielded a primary study of 1,284 respondents, each completing two of the following six types of importance measurement, and in a random order:



- Standard importance ratings
- “Unbounded ratings”³
- Magnitude estimation⁴
- Constant sum
- Q-sort
- Maximum difference scaling

Examples of the specific formats for these methods appear in the aforementioned *International Journal of Market Research* article.

In each task, respondents reported the importance of 10 attributes known to affect preference for casual dining restaurants:

- Prompt greeting
- Overall cleanliness
- Comfortable environment
- Server attentiveness
- Server friendliness
- Pace of meal
- Taste of food
- Temperature of food
- Check arrives in timely manner
- Reasonable prices

In addition, each respondent rated the casual dining restaurant they visited most recently both overall and in terms of each of the 10 attributes.

Results of Empirical Tests

Time to Complete Task – on average, the importance measures took about a minute and a half. Two exceptions were stated importance ratings, which took about half as long to complete, and maximum difference scaling, which took about twice as long:

Method	Median seconds to complete
Importance ratings	38
Q-sort	75
Unbounded ratings	79
Magnitude estimation	82
Constant sum	89
Maximum difference scaling	171

So in terms of questionnaire real estate, maximum difference scaling is a bit of a space hog, adding 90 seconds to the questionnaire more than every other method.

Discriminating Power – we can test for the size of differences between attributes’ importances for each of the methods using a statistical test, “repeated measures analysis of variance.” Resulting from these analyses are F statistics, where a larger F means more differentiation between attributes and a smaller F means less.

Method	F
Maximum difference scaling	4,031
Q-sort	484
Constant Sum	131
Importance ratings	115
Magnitude estimation	66
Unbounded ratings	17

The F statistics suggest that all the methods detect significant differences between attributes (at $P < .01$), but recall this study included a large sample size. In studies with sample sizes more typical of applied marketing research, methods on the lower half of the table may not produce significant differences at all. Two methods, maximum difference scaling and Q-sort clearly stand out as performing much better than the other measures while magnitude estimation and unbounded ratings do especially poorly.

Predictive power – remember the multi-attribute attitude model mentioned in the Introduction? It turns out this model makes the perfect vehicle for comparing the predictive power of the six stated importance measures.

For starters, we can use the respondents’ performance ratings of the 10 attributes to predict their overall satisfaction with their most recent casual dining experience. When we run this regression analysis we get a correlation of actual with predicted satisfaction of 0.31.

When we include each of the importance measures in the model in turn, the correlation between actual and predicted satisfaction will improve if a given importance measure has predictive validity and not if it lacks predictive validity.

Method	Correlation
Maximum difference scaling	0.62
Q-sort	0.61
Constant Sum	0.60
Magnitude estimation	0.55
Importance ratings	0.30
Unbounded ratings	0.26

As academics found with the multi-attribute attitude model in the 1970s, standard importance ratings not only fail to raise the fit between actual and predicted satisfaction, they actually reduce it. As bad as that is, unbounded ratings manage to reduce it even further! These two methods lack predictive validity, making them utterly worthless as measures of attribute importance at the individual respondent level.¹

Summary

Maximum difference scaling has the greatest discriminating and predictive power, but also takes the longest of any of the methods for respondents to complete. Add the experimental design work and the data pre-processing required to get importances from maximum difference scaling, and Q-sort starts to look like the better option for both applications. Both of these methods work well in web-based surveys and in-person surveys, while maximum difference scaling can also work in a mail survey. A phone survey will support neither of these methods.

Standard importance ratings cost little in terms of respondent effort, which may be why they fare well on predictive power and fare poorly on discriminating power.

Unbounded ratings did poorly both in terms of predictive validity (they had none) and in terms of discriminating

power (very little), making it even worse than standard importance ratings.

Constant sum and magnitude estimation perform almost as well as the best methods in terms of predictive power, though constant sum outperforms magnitude estimation when it comes to discriminating power. Magnitude estimation has the benefit of being the best-performing method suitable for telephone interviewing.

Maritz Research recommends:

- Maximum difference scaling or Q-sort for web-based or in-person surveys
- Maximum difference scaling or constant sum for mail surveys
- Magnitude estimation for phone surveys

Footnotes

¹ Bass, F.M. & Wilkie, W.L. (1973) "A comparative analysis of attitudinal predictions of brand preference," *Journal of Marketing Research*, **10**, pp. 262-9. Beckwith, N.E. & Lehmann, D. (1973) "The importance of differential weights in multi-attribute models of consumer attitude,"

Journal of Marketing Research, **10**, pp. 141-5. Wilkie, W. L. and Pessemier, E. A. (1973) "Issues in marketing's use of multi-attribute attitude models," *Journal of Marketing Research*, **10**, pp. 428-41.

² Chrzan, K. and N. Golovashkina (2006) "An empirical test of six stated importance measures," *International Journal of Market Research*, **48**, 717-40.

³ Marder, E. (1997) *The Laws of Choice: Predicting Consumer Behavior*. New York: Free Press.

⁴ Lodge, M. (1981) *Magnitude scaling: Quantitative measurement of opinions*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-025 Beverly Hills: Sage Publications

www.maritzresearch.com
(877) 4 MARITZ
info@maritz.com

