An empirical test of six stated importance measures

Keith Chrzan and **Natalia Golovashkina**
Maritz Research

This paper reports on a web-based commercial customer satisfaction study consisting of 1284 respondents, which measured stated attribute importance using six different methods (importance ratings, constant sum, Q-sort, maximum difference scaling, unbounded ratings and magnitude estimation). Statistical analyses were used to evaluate these six methods in terms of (a) the time they take to administer, (b) their ability to provide discriminating measures and (c) their predictive validity. Clear winners and losers emerge from these analyses, and applied marketing researchers can use these findings to the benefit of the marketers they support.

Background

In order to sell their products or services, marketers, brand managers and new product development teams need to know which aspects of their products or services are more important than others. By measuring, monitoring and enhancing performance on the important product or service attributes, marketers seek to increase the value of their offerings to the market and thus to increase sales.

To support these efforts, applied marketing researchers measure attribute importance in a variety of research situations. For example:

- in customer satisfaction studies, to identify the aspects of the customer experience that most affect customer satisfaction and loyalty
- in brand preference research, to identify the attributes that most affect market share
- in segmentation research, to identify 'needs-based' segments of customers who can be attracted by different value propositions

Received (in revised form): 4 April 2006

- in new product development research, to identify aspects of a potential product that can be changed to improve its chances of success.

Researchers have at their disposal at least three classes of tools for measuring importance. First, they can simply ask respondents for direct evaluations of attributes in terms of their importance, a strategy known as 'stated' importance measurement. In 'derived' importance measurement, on the other hand, researchers quantify importance indirectly, using statistical analysis to tease out importance from the relations between some sort of overall evaluation of a product or service and a set of performance measures that describe how much the product or service possesses the various attributes. Finally, the researcher can take even more control of the measurement situation and give respondents an experimentally designed set of stimuli that differ systematically on the various attributes in ways known to the researcher. Statistical analysis of these controlled stimuli results in conjoint analysis, yet another way of measuring attribute importance.

Conjoint methods have emerged as the standard for new product development research and a staple of academic journals and conferences. A number of conjoint methods have evolved and they continue to be refined, compared and tested. Derived importance models feature prominently in many studies of customer satisfaction (Myers 1999; Allen & Rao 2000), of brand preference (Myers 1996) and of brand choice (McFadden 1973; Gensch & Recker 1979).

Some widely recognised shortcomings (Bass & Wilkie 1973; Myers 1999; Allen & Rao 2000) of one form of stated importance measure, namely importance ratings, have limited the use of stated importance methods: some practitioners use them in satisfaction and brand preference studies, but their most consistent use is probably in providing measurement inputs for needs-based segmentation studies. Academic interest in stated importance measures waned after the protracted demise and burial of the multi-attribute attitude model (Fishbein 1967; Bass & Wilkie 1973; Beckwith & Lehmann 1973; Wilkie & Pessemier 1973) in the 1970s.

Our interest as applied marketing researchers is to identify viable methods of stated attribute importance. To do this we subject several methods to relevant and discriminating empirical tests. In this study we are extending the work that Orme (2003) and Cohen (2003) performed in comparing stated importance ratings to maximum difference scaling and to a hierarchical Bayesian version of the method of paired comparisons.

Methods available to practitioners

Importance ratings

Attribute importance ratings are easy to collect: they take up little questionnaire real estate, and respondents find them familiar and easy to answer. Unfortunately, researchers and practitioners have noted two obvious problems with importance ratings: (1) they may be affected by social desirability bias (Bacon 2003); and (2) in the absence of constraints, or any need to make trade-offs, respondents can (and often do) give all attributes high importance ratings, so that the responses for all attributes tend to bunch up towards the positive end of the rating scale (Myers 1999). This makes it hard to tell more from less important attributes and it can impair subsequent multivariate analyses that rely on discrimination among the importance ratings, such as needs-based segmentation.

An even more damning flaw of stated importance ratings is their evident lack of predictive validity. A variety of researchers tried to use importance ratings as part of the compositional 'multi-attribute attitude model' (Rosenberg 1956; Fishbein 1967). This model posits that the sum of the products of attribute importances and a brand's performance on those attributes should be a measure of the overall appeal of that brand. More formally,

$$O \approx \sum_{i=1}^{i=k} I_i P_i$$

where O is an overall evaluation of the brand, I_i is the importance of the ith attribute and P_i is the performance of the brand on the ith attribute, and the summation occurs over K attributes.

When researchers built the multi-attribute attitude model using importance ratings as measures of importance, they pretty routinely found that the model was no better (and often worse) when importance ratings were included than when they were left out (Bass & Wilkie 1973; Wilkie & Pessemier 1973).

Alternatives to importance ratings

Although importance ratings scales remain in use today, practitioners also have other methods available to them. Magnitude estimation (Lodge 1981) gives a respondent one attribute as an anchor, and asks them to rate all other attributes in relation to it. Proponents of magnitude estimation

suggest that ratio measures of attribute importance result, but to anyone who has worked with survey research data, it may seem doubtful that respondents can perform the calculations required for this kind of scaling. In particular, respondents may more easily compute multiples than quotients. The sample questionnaire in the Appendix shows how one might pose the magnitude estimation task to respondents.

Eric Marder (1997) suggests an innovative 'unbounded' rating scale as an alternative to conjoint measurement. Although Marder does not directly suggest such a scale for importance measurement, the stretch is not a great one, and a modified unbounded scale merits attention. The unbounded rating scale has received scant academic attention (Stapleton & Edmonds 2005) and produced mixed results, with the authors suggesting that the scale may be more useful for within-respondent comparisons than for across-respondent comparisons.

Both of these methods take the tack of removing constraints from the way respondents answer importance questions. Both allow respondents much greater freedom in how they respond to importance stimuli than do importance ratings.

Other methods do the opposite: they impose structural constraints on the answers respondents provide, forcing them to make trade-offs. For example, constant sum scaling prevents respondents from giving high ratings to all attributes, because it requires them to make trade-offs – they cannot allocate a large number of points to one attribute without taking them away from others. Constant sum measurement forces trade-offs, but still gives respondents plenty of latitude regarding how to distribute their responses. Myers (1999) notes that constant sum scaling may not work well with larger numbers' (>10) attributes.

The remaining methods force trade-offs by imposing strict constraints even on the distribution of respondents' answers to importance questions. The simplest of these requires respondents to rank attributes in terms of their importance, in effect imposing a uniform distribution on responses. Respondents find ranking large numbers of attributes difficult, however, and most practitioners shy away from collecting importance rankings.

The Q-sort methodology (Stephenson 1953) imposes a quasi-normal distribution on responses, rather than a uniform distribution. For example, in a Q-sort task involving 25 items, a researcher may direct respondents to score the items into seven categories, as follows:

1. 1 into the category denoting greatest importance
2. 2 into the next most important category

3. 5 into the third category
4. 9 into the middle category
5. 5 into the slightly unimportant category
6. 2 into the next to the least important category, and
7. 1 in the least important category.

The quasi-normal 1:2:5:9:5:2:1 distribution of responses is fixed for all respondents: they are unable either to rate all attributes as highly important or as unimportant, so the method imposes severe trade-offs. See the Appendix for an example of how to implement a Q-sort task in a web-based survey.

The method of paired comparisons, or MPC, traces back to Thurstone (1927), with important refinements from Bradley and Terry (1952) and David (1988). Respondents receive pairs of attributes and indicate which member of the pair they find more important. Each respondent answers several such pairs. MPC formerly provided only aggregate measures of importance, but the advent of hierarchical Bayesian (HB) analysis (Orme 2003; Retzer 2006) provides a parsimonious way to use paired comparison stimuli to create respondent-level importance data, provided respondents evaluate 1.5 times as many pairs as there are attributes.

An interesting multiple-choice extension of the method of paired comparisons is 'best/worst' or maximum difference (maxdiff) scaling (Finn & Louviere 1992). In a typical maxdiff scaling question a respondent may see four to six attributes, among which he or she would specify the most important and the least important, as follows:

> Which of the following aspects of a visit to a casual dining restaurant is *most* important to you and which is *least* important to you?

Aspect	**Most**	**Least**
Prompt greeting	[ ]	[ ]
Server attentiveness	[ ]	[ ]
Taste of food	[ ]	[ ]
Reasonable prices	[ ]	[ ]

As with MPC, respondents evaluate several such sets (often only as many sets as there are attributes) and HB multinomial logit produces importance estimates for each attribute for each respondent.

Previous comparative studies

Two recent studies tested some of these methods against one another. Orme (2003) compared importance ratings to respondent-level MPC measures of attribute importance and found the MPC estimates to have greater predictive validity than importance ratings, better ability to discriminate importance of the attributes themselves, and better ability to show differences in attribute importance across a priori segments of respondents. Using the same data sets, Cohen (2003) found maxdiff estimates to outperform attribute ratings and (slightly) to outperform MPC importances on the same measures (predictive validity, ability to discriminate).

Current empirical study

Research design

We seek to extend the work begun by Orme (2003) and Cohen (2003) to include more of the alternative methods for stated attribute importance. Building upon their work, we include in our test the best method from their studies, namely maxdiff scaling, and the worst, namely importance ratings, the de facto industry standard.

In addition our test includes four of the other stated importance measurement methods described above:

1. magnitude estimation
2. unbounded ratings
3. constant sum
4. Q-sort.

We included this test as part of a commercial study of customers' satisfaction with casual dining restaurants. A total of 1284 casual dining customers responded to an email invitation, passed screening questions and completed a web-based survey. Surveys averaged about 8 minutes in length and all were completed between 3 and 7 January 2006.

Respondents were members of the Esearch internet panel, which we have found to be broadly representative of the United States internet population. Screened to have eaten at a casual dining restaurant at least once in the 30 days preceding the survey, respondents reported an average of 4.4 (in the preceding 30 days) casual dining occasions. Brand shares track well with market data and respondent demographics are not

obviously biased in representing the US casual dining market: 48% male, 45% under age 50, 38% with university degrees, 73% married or coupled, 37% with children in the household and 54% with household incomes in excess of $50,000 per year.

Researchers experienced in the study of casual dining restaurants together with a restaurant marketing industry expert came up with the list of ten attributes they believe drive satisfaction with the casual dining experience. Respondents rated their overall satisfaction with one of five brands of casual dining restaurant, and they rated the performance of that brand on each of the ten attributes.

In addition, each respondent completed a randomly selected two of the six importance measurement methods, in a random order – so each method was as often the second importance measure collected as the first. Thus about 428 respondents completed each of the six stated importance batteries, as follows:

1. 428 importance ratings
2. 428 constant sum
3. 429 maximum difference scaling
4. 429 Q-sort
5. 428 unbounded ratings
6. 426 magnitude estimation.

Order effects were slight, but one was interesting: standard importance ratings fared better in terms of predictive validity when they were the second task a respondent completed than when they were first.

Questionnaire administration and exploratory data analysis

Exact wording of the different questions appears in the Appendix, but brief descriptions and results of exploratory data analyses follow. Table 1 contains mean importance scores for the ten attributes and six stated importance methods.

All six methods identify the same two most important attributes: Taste of food and Overall cleanliness. All methods except attribute ratings identify Reasonable prices as the third most important attribute. All methods except constant sum identify the same two least important attributes: Prompt greeting and Receive bill in timely manner.

Previous researchers (Heeler *et al.* 1979; Jaccard *et al.* 1986) noted a lack of convergent validity among various methods of measuring stated

Table 1 Mean importances, by method

Attribute	R	CS	QS	MD	UR	ME
Prompt greeting	2.8	4.2	1.9	0.9	28.4	25.8
Overall cleanliness	4.2	14.6	3.6	15.4	73.6	56.3
Comfortable environment	3.5	7.9	3.2	4.3	54.1	40.6
Server attentiveness	3.5	8.9	3.1	5.5	53.0	42.2
Server friendliness	3.5	7.9	2.9	6.2	53.7	40.3
Pace of meal	2.9	4.0	2.2	1.0	27.6	27.5
Taste of food	4.3	23.0	4.3	34.5	82.2	66.3
Temperature of food	3.8	10.7	3.1	12.5	59.4	45.4
Receive bill in timely manner	2.9	4.7	2.0	0.8	30.1	26.5
Reasonable prices	3.7	14.0	3.7	18.9	64.4	50.0

R = standard importance ratings; CS = constant sum scaling; QS = Q-sort;
MD = maxdiff scaling; UR = unbounded ratings; ME = magnitude estimation

Notes: All tables use importance measures transformed as described in the text:

1. Maxdiff utilities were exponentiated and converted via the logit choice rule transformation to sum to 100%.
2. Magnitude estimates were log transformed for analyses and prior to taking the mean. For reporting the mean we exponentiated back to the original scale; this method (recommended by Lodge 1981) reduces the effect of outliers on mean scores.
3. Unbounded ratings were transformed as described in the text, dividing each by the maximum of the importances to account for the fact that different respondents had different anchors on their scales.

importance, but these six methods produce results largely in agreement. Correlations among the vectors of the ten attributes' mean importances on the six methods are all high, none of them being lower than 0.87 (Table 2).

It could be that the nature of attributes in this category for some reason promotes greater convergent validity than do the attributes in the categories studied by the previous authors noted above. Perhaps having additional evidence from more product categories will shed light on this issue.

Table 2 Convergence of methods

Method	R	CS	QS	MD	UR	ME
Ratings	1.00					
Constant sum	0.93	1.00				
Q-sort	0.95	0.93	1.00			
Maxdiff	0.87	0.99	0.90	1.00		
Unbounded ratings	0.99	0.93	0.98	0.88	1.00	
Magnitude estimation	0.99	0.97	0.98	0.93	0.99	1.00

See Table 1 for explanation of abbreviations.

Importance ratings
For importance ratings we used a five-point fully anchored scale that we have found in previous research to spread out responses more than do other fully anchored importance scales, and more than does an endpoint-only anchored scale. The anchors for the five scale points were:

1. Not at all important
2. Somewhat important
3. Very important
4. Extremely important
5. Critically important.

We again found with this scale that the extreme wording on the top scale point may have inclined respondents not to bunch up on the top two scale points as readily as with other importance rating scales. Across respondents and across the ten attributes, the distribution of importance ratings was:

2.1%	Not at all important
16.8%	Somewhat important
30.0%	Very important
29.8%	Extremely important
21.2%	Critically important

Attribute order we randomised across respondents. Thirty-three respondents (7.7%) gave the same rating to all attributes, an entirely uninformative pattern that does not allow any discrimination of importance.

Constant sum
Respondents allocated 100 points across the ten attributes in terms of their importance. Logic built into the web-based survey prevented responses from exceeding 100% in sum and from being negative. Again we randomised attribute order across respondents. Thirty-four respondents (7.9%) split the 100 points evenly among the ten attributes, assigning ten points each, for no variance in importance weights. At the other extreme, only one respondent assigned all 100 points to a single attribute.

Maximum difference scaling
Respondents completed ten maxdiff questions, each with four of the attributes. Questions conformed to a well-balanced experimental design,

and each item appeared exactly four times across the ten questions. Although we did not randomise attribute order across respondents, we did control for attribute position in the experimental design, ensuring that each attribute appeared once each in the first, second, third and fourth position in its four appearances. Because of the strongly constrained response format, respondents would have had great difficulty giving the same importance for all attributes (in fact, none of them did).

Q-sort

Again, attributes appeared in a random order across respondents. We operationalised Q-sort by having respondents first choose their most important attribute from the list of ten, which was coded as '5' in the data file, as it was removed from the screen. Respondents then chose their next two most important attributes from the nine remaining, and these we coded as '4's and removed from the screen. From the seven remaining attributes respondents next identified the least important (coded '1') and we removed it, leaving six. Finally, the respondents specified two as the next least important attributes (coded '2' in the data file). We coded the four unchosen attributes into category 3. Again, because of the strongly constrained response format, respondents were unable to assign equal importance to all attributes.

Unbounded rating scale

We instructed respondents to write in as many Us (for 'unimportant' or Is (for 'important') as they felt represented their opinions about each attribute. Respondents used '0' to connote indifference towards an attribute. While some respondents never wrote in more than a single I or U, others wrote in as many as 213 Is and up to 54 Us. Forty-three of the respondents (10.2% of those completing the unbounded ratings) gave the same response for all attributes and 40 of these gave all attributes a score of 1. The respondent providing the largest range between maximum and minimum scores had a range of 166.

Magnitude estimation

One attribute, randomised to differ across respondents, served as the anchor for each respondent, with a value of 50. Respondents then rated the other nine attributes using their reference attribute and its importance of 50. Nine respondents (2.1%) reported the same importance for each attribute, and the greatest range between a respondent's high and low scores was 99,950. Unfortunately, when we standardise the responses to

account for respondents' different reference attributes (dividing each attribute's value by that of the tenth attribute, and then multiplying by 50) we find that respondents' different starting-points cause significant differences on some of the attributes. This sensitivity to context effects may have hampered the performance of the magnitude estimation scale in our analyses below. Of course it also calls into doubt the validity of the magnitude estimation scale.

Uninformative responses

In four of the methods (all but Q-sort and maxdiff scaling), respondents could choose to give all attributes equal importance, and these uniform evaluations represent uninformative responses. Interestingly, respondents who gave uninformative responses in one task did not necessarily do so in the other. For example, of the 86 respondents who completed both importance ratings and constant sum, 75 gave informative answers to both types of questions, four gave uninformative responses to both, three gave uninformative responses only to importance ratings and four gave uninformative responses to the constant sum. Similar results occurred for the other pairings or measures.

Data pretreatment

Rating scales, constant sum scores and Q-sort measures result directly from responses to survey questions. The other three measures require some analysis and/or scaling before use in the comparative analyses below.

Because we randomise across respondents which attribute serves as the 50-point anchor for the magnitude estimation task, we correct for this by normalising all respondents' scores, dividing all of them by the score for the tenth attribute. Moreover, Lodge (1981) notes the impact that outliers can have on the magnitude estimation method. In our study, a handful of respondents gave importances of 1000 or more relative to an anchor of 50, and one response went as high as 100,000. Lodge recommends using the log rather than the raw responses, advice we follow in the analyses below (all of which worked better with log-transformed importances than with the untransformed scores).

The unbounded scale shows strong between-respondent variations in scale usage. Many respondents confine themselves to a range of –1 to +1, but others use a range of well over 100 points. This artifact we correct for by dividing all ten importances for each respondent by the absolute value of the highest score for that respondent.

Finally, the maximum difference scaling task requires hierarchical Bayesian multinomial logit analysis to produce its respondent-level importance estimates. Scale differences occur among respondents because of different degrees of the fit of their data to their utility model (Swait & Louviere 1993). These we remedy by using the MNL choice rule to express each utility as a share of preference for its attribute relative to other attributes:

$$Preference = \frac{\exp(utility_k)}{\sum_{i=1}^{i=10} \exp(utility_i)}$$

Like constant sum, this transformation results in a set of importances that are ratio level and that sum to 100%.

Planned analyses

Task length

Questionnaire real estate costs money, so we measured the length of time it took respondents to complete each importance measurement task. Because respondents could pause while taking the survey, a small number of very long task times result. These outliers inflate the variance of the times and render parametric tests like the *t*-test unreliable. To combat this we test median times rather than mean times, using a non-parametric χ^2 test of medians.

Between-item discrimination

Ideally an importance measurement technique will be discriminating: it should allow a researcher to distinguish more from less important attributes. For each method, we could test the 45 unique pairs among the ten attributes, and count the number of significant results. A better omnibus test, however, is a repeated measures analysis of variance, where each attribute is a repeated observation; this produces an *F*-statistic we can compare across methods. Because of the normalisation to the magnitude estimation described above, the tenth attribute is a constant value of 50 and we include only the 36 pairs of the other nine attributes.

Between-group discrimination

The survey includes six demographic questions on which respondents might be split into a priori groups or segments. Using these to create nearly

even splits, we can then test how many of the 60 (six grouping splits times ten attribute importances, but only 54 for magnitude estimation) resulting *t*-tests are significant for each of the attribute importance measures. Clearly the more of these that are significant, the more discriminating power the importance measurement method possesses. We also report the average *t*-statistic for each of the six attribute importance measurement methods.

Predictive validity

We employ two comparisons of predictive validity, one aggregate and one disaggregate. The aggregate measure of predictive validity compares the results of the six stated importance methods to a derived importance model. The survey included an overall satisfaction rating for one of the brands of casual dining restaurant and performance ratings for the ten attributes of that brand. We can model the overall rating as a function of the attribute performance ratings and the resulting coefficients are derived importance ratings.

Though in common use, both correlations and regression coefficients make poor derived importance measures. The former ignore multivariate relations among the predictors and, because the predictors often suffer from a halo effect (Beckwith & Lehmann 1973), they tend all to be highly correlated with the overall variables. The same problem impacts regression analysis by creating multicollinearity. Multicollinearity causes well-recognised, very serious and entirely pervasive problems in derived importance analysis, particularly in customer satisfaction studies (Strathmann *et al*. 1994; Myers 1999; Allen & Rao 2000). Regression analysis in the presence of multicollinearity can produce highly misleading results, even reversing the signs of some coefficients (suggesting, for instance, that reducing quality, providing abusive service or raising prices will increase satisfaction with a brand).

The derived importance model we use avoids the shortcomings of both correlation analysis and regression analysis; we employ Theil's suggestion to supplement Kruskal's averaging over orderings strategy with an information measure (entropy reduction) rather than partial R^2 statistics (Kruskal 1987; Theil & Chung 1988; Soofi *et al*. 2000). This model expresses attribute importances as percentages of information content (similar to percentages of explained variance attributable to each attribute). These importances result from the True Driver Analysis (Table 3).

The model's R^2 of 0.36 implies that the correlation between actual and predicted overall satisfaction is 0.60.

Using this more defensible measure of derived importance, we can look at correlations of each of the vectors of mean importances with the derived importances. Each vector will have ten elements, one for each of the attributes. This aggregate analysis is descriptive and suggestive, but not statistically powerful: with only ten observations, significant differences in correlations will be difficult to detect.

Table 3 Mean importances, by method

Attribute	Derived importance
Prompt greeting	5.7
Overall cleanliness	5.1
Comfortable environment	5.5
Server attentiveness	14.6
Server friendliness	7.4
Pace of meal	8.5
Taste of food	30.5
Temperature of food	8.4
Receive bill in timely manner	5.3
Reasonable prices	9.1

A disaggregate test of predictive validity, however, allows us to correlate predicted and actual results at the individual respondent level and thus provides much more statistical power. For this test we resuscitate the multi-attribute attitudinal model described above – multiplying each respondent's stated importance of each attribute by their ratings of that attribute's performance, and then summing these ten products to yield a prediction of the overall appeal of the brand to each respondent. Across individuals this yields the correlation of actual appeal and overall ratings based on about 428 cases for each of the six stated importance methods. While some measures are arguably ratio in nature (e.g. constant sum scales), we know that some are only interval scales, at best (e.g. rating scales). To avoid unfairly penalising non-ratio-level importance measures, we will use the optimal scaling corrections suggested by Holbrook (1977).

Results

Task length

Median task lengths for each of the six stated importance methods are shown in Table 4.

Table 4 Median length of importance measurement methods, in seconds

Method	Median length (seconds)
Ratings	38.0
Q-sort	75.0
Unbounded ratings	79.0
Magnitude scaling	82.0
Constant sum	88.5
Maxdiff scaling	171.0

The omnibus χ^2 test is significant at $p < 0.001$ (χ^2 of 724.86 with 5 df), as are most of the pairwise follow-up tests, as differences of 7 seconds or greater are significant.

The median respondent takes less than 4 seconds on each attribute rating question.

At the other extreme, the maxdiff task takes respondents nearly 3 minutes to complete, an investment in survey real estate that could be critical in some applications with greater numbers of attributes.

Between-item discrimination

Omnibus *F*-statistics for the repeated measures analysis of variance of each of the six stated importance methods all show significant differences among the attributes' importances ($p < 0.001$) – see Table 5.

Table 5 *F*-statistic for between-item discrimination, by method

Method	*F*
Ratings	209
Constant sum	166
Q-sort	357
Maxdiff	578
Unbounded ratings	125
Magnitude estimation	123

The larger *F*-values for Q-sort and particularly for maxdiff show them to be the more highly discriminating measures. This makes them liable to work well even in situations with smaller sample sizes, and hence less statistical power, than in the current study. Since discriminating more from less important attributes is the objective of collecting importance measures, this greater power to discriminate is crucial.

Between-group discrimination

Testing the ten attributes (nine for magnitude estimation) across splits in seven a priori segmentation variables produces 60 *t*-tests (50 for magnitude estimation) for each type of stated importance measurement. The second column of Table 6 shows how many of the *t*-tests are significant, and the third column shows the average *t*-value across the 60 tests for each method.

All of the methods produce more differences than one would expect from chance alone. Although Q-sort and constant sum show the most differences, average *t*-statistics for all methods are small, certainly smaller than in the studies reported by Orme (2003) and Cohen (2003).

Table 6 *t*-tests for between-group discrimination, by method

Method	No. significant	Average *t*
Ratings	9	1.11
Constant sum	12	1.24
Q-Sort	14	1.24
Maxdiff scaling	10	1.23
Unbounded ratings	8	0.92
Magnitude scaling	9	1.32

Predictive validity

The first analysis of predictive validity looks at the correlation of each of the methods with derived importance. These correlations are based on just the ten summary attribute importances for each method, and statistical tests for differences among them are correspondingly weak and non-significant (Table 7).

Table 7 Correlations of stated importances with derived importances

Method	Correlation
Ratings	0.56
Constant sum	0.75
Q-Sort	0.63
Maxdiff scaling	0.77
Unbounded ratings	0.55
Magnitude scaling	0.63

Maxdiff scaling and the constant sum methods produce mean importances that most resemble the vector of derived importances from the True Driver Analysis, possibly because they, like importances from True Driver Analysis, sum to 100%.

The second analysis of predictive validity uses respondent-level data and looks at the correlation between overall satisfaction with a brand and the prediction of overall performance based on the multi-attribute attitude model (Table 8).

Three of the methods, namely constant sum, Q-sort and maxdiff, have correlations with the actual overall satisfaction measure that equal the 0.60 correlation produced by the best-fitting derived importance model. On this powerful measure of predictive validity, these three measures perform the best, closely followed by magnitude scaling. Traditional importance ratings and unbounded importance ratings both perform poorly. Indeed, their inferior predictive validity exemplifies the poor performance which contributed to the abandonment of the multi-attribute attitude model nearly 30 years ago – both of these actually make the multi-attribute attitude model worse by their inclusion!

As expected, optimal scaling improved the fit of the multi-attribute attitude model for rating scales, and it has almost no effect on constant sum and maxdiff scaling, which look like ratio-level data, to begin with. We find it surprising that Q-sort results improve very little through optimal scaling, suggesting that their highly constrained response format does not

Table 8 Correlations of multi-attribute attitude model predictions with overall satisfaction

Method	Correlation	After optimal scaling
Ratings	0.30	0.48
Constant sum	0.60	0.60
Q-Sort	0.61	0.61
Maxdiff scaling	0.62	0.63
Unbounded ratings	0.26	0.50
Magnitude scaling	0.55	0.62

prevent respondents from using the Q-sort to express their opinions. In addition, the unbounded ratings improve the most underoptimal scaling, confirming the peculiarity of their original scale and suggesting they be used with caution, if at all.

Conclusions

Attribute importance ratings are easy for respondents, who do them very quickly. They are also, however, together with unbounded ratings, the worst performing of the six importance measures we tested.

Maxdiff scaling performs the best in terms of ability to differentiate importance among the attributes and it has the highest predictive validity. A drawback is that it takes the longest of the six methods to administer, but if measuring attribute importance contributes prominently to a study's objectives, it may be an investment in questionnaire real estate worth making.

Q-sort has the advantage of taking less time to administer than any method other than importance ratings, and of producing results, in terms of discrimination and prediction, very nearly as good as maxdiff. Although we have found Q-sorts of long lists of attributes difficult to administer in web-based surveys, it is an attractive method for many studies with shorter attribute lists, or for studies administered in person.

Constant sum and magnitude estimation had lower discriminating power than maxdiff or Q-sort, but similar predictive validity. In the bygone days of the multi-attribute attitude model, these methods would have seemed promising, but the work that model attempted has become the rightful domain of conjoint methods. Because they are not as discriminating as Q-sort and maxdiff, and because between-attribute discrimination will make contemporary uses of stated importance measures useful, we see limited applications for these methods.

Discussion

In a study with large numbers of attributes, Q-sort may be difficult to implement, leaving maxdiff as the method of choice. Maxdiff, Q-sort and constant sum all work best when respondents can see the survey questions – that is, in web-based, mail or in person research modalities. If one must collect importance information by phone, magnitude estimation beats the two worst-performing methods, namely ratings and unbounded ratings.

As noted above, our study used real-world respondents in a commercial marketing research study, using typical incentives and data-collection

methods. That they were not college students in an artificial setting should ameliorate some concerns about the generality of the findings.

Of course we have used just a single product category in this study. Whether its findings generalise to other populations, other cultures or other product categories, and whether the findings hold across different kinds of data-collection methods, we leave to future creative researchers who want to extend and improve our testing.

Appendix: Question wording

Of course the online survey had the visually appealing layout and professional formatting typical of a web-based questionnaire. Programming was performed by the web-survey experts at Critical Mix.com, who also hosted the survey. A text approximation of the importance measurement sections of the questionnaire is as follows (programming notes shown IN CAPITALS).

Order and identity of methods

THERE ARE 6 VERSIONS OF QUESTION 4. EACH RESPONDENT GETS JUST 2 OF THESE 6. RANDOMISE WHICH 2 VERSIONS EACH RESPONDENT GETS. RANDOMISE THE ORDER IN WHICH RESPONDENTS GET THE TWO VERSIONS. THERE ARE 30 QUOTA CELLS – FOR EACH OF THE 15 PAIRS OF VERSIONS OF QUESTION 4 AND FOR WHICH OF THE TWO VERSIONS SUMS FIRST AND WHICH SECOND. PLEASE INSERT TIMERS SO THAT WE KNOW HOW MANY SECONDS EACH OF 4A–4F TAKES RESPONDENTS TO COMPLETE.

Importance ratings

4a. Please indicate how important each of these aspects of a casual dining restaurant is to you. Use the scale across the top of this list to describe how important each aspect is to you. RANDOMISE ORDER.

	Not at all important	Somewhat important	Very important	Extremely important	Critically important
Prompt greeting	[ ]	[ ]	[ ]	[ ]	[ ]
Overall cleanliness	[ ]	[ ]	[ ]	[ ]	[ ]
Comfortable environment	[ ]	[ ]	[ ]	[ ]	[ ]
Server attentiveness	[ ]	[ ]	[ ]	[ ]	[ ]
Server friendliness	[ ]	[ ]	[ ]	[ ]	[ ]
Pace of meal	[ ]	[ ]	[ ]	[ ]	[ ]
Taste of food	[ ]	[ ]	[ ]	[ ]	[ ]
Temperature of food	[ ]	[ ]	[ ]	[ ]	[ ]
Receive bill in timely manner	[ ]	[ ]	[ ]	[ ]	[ ]
Reasonable prices	[ ]	[ ]	[ ]	[ ]	[ ]

Constant sum

4b. Please indicate how important each of these aspects of a casual dining restaurant is to you. To do this, please distribute 100 points across these aspects according to how important each is to you. You can give as many or as few points as you like to each aspect, but the total number of points you assign must be 100. RANDOMISE ORDER, MAKE SURE TOTAL SUMS TO 100.

Prompt greeting	_____
Overall cleanliness	_____
Comfortable environment	_____
Server attentiveness	_____
Server friendliness	_____
Pace of meal	_____
Taste of food	_____
Temperature of food	_____
Receive bill in timely manner	_____
Reasonable prices	_____
Total	**100**

Maximum difference scaling

4c. Please indicate how important each of these aspects of a casual dining restaurant is to you. In each of the following sets of items, please pick the ONE item you consider MOST important and the ONE item you consider least important. SEE MAXDIFF.XLS FOR A FORMATTED WORKSHEET FOR THIS SET OF QUESTIONS, 4C1 TO 4C10.

Q-sort

4d. Please indicate how important each of these aspects of a casual dining restaurant is to you. Which of the following aspects is MOST important to you? REQUIRE 1 CHOICE. RANDOMISE ORDER, BUT ONLY AT THE START, FOR THE FIRST QUESTION – FROM THERE KEEP ORDER CONSTANT.

Prompt greeting	_____
Overall cleanliness	_____
Comfortable environment	_____
Server attentiveness	_____
Server friendliness	_____
Pace of meal	_____

Taste of food _____
Temperature of food _____
Receive bill in timely manner _____
Reasonable prices _____

REMOVE THE ONE CHOSEN AND ASK Which **two** of the remaining aspects of casual dining restaurants are the **most** important to you? Please pick 2. REQUIRE 2 CHOICES.

REMOVE THE TWO CHOSEN AND ASK Now let's change direction and think about the **least** important things. Which one of these aspects is the **least** important to you? REQUIRE 1 CHOICE.

REMOVE THE ONE CHOSEN AND ASK Finally, which **two** of the remaining aspects of a casual dining restaurant are the **least** important to you? Please pick 2. REQUIRE 2 CHOICES.

PLEASE CODE THE ONE MOST AS '5', THE TWO NEXT MOST AS '4', THE 4 UNCHOSEN AS '3', TWO NEXT LEAST AS '2' AND THE VERY LEAST AS '1'.

IF CRITICAL MIX HAS A BETTER WAY OF ASKING OR SHOWING Q-SORT QUESTIONS LIKE THIS ONE, PLEASE LET ME KNOW.

Unbounded scale

4e. Please indicate how important each of these aspects of a casual dining restaurant is to you. For each of the following aspects of a casual dining restaurant, type in the letter 'I' as many times as you like to indicate how important it is to you – the more important it is to you, the more Is you should type next to it. If an aspect is unimportant to you, type in U or UU or UUU or as many Us as you want – the more unimportant it is to you, the more Us you should type next to it. Please don't leave any box blank: if you are neutral about an aspect, just type a '0' in the box. RANDOMISE ORDER.

Prompt greeting _____
Overall cleanliness _____
Comfortable environment _____
Server attentiveness _____
Server friendliness _____
Pace of meal _____
Taste of food _____
Temperature of food _____
Receive bill in timely manner _____
Reasonable prices _____

Magnitude estimation

FOR THE NEXT QUESTION, RANDOMLY CHOOSE ONE OF THE 10 ASPECTS WE'VE BEEN USING (THE LIST OF 'PROMPT GREETING' TO 'REASONABLE PRICES'). RECORD THE ONE CHOSEN, WHICH BECOMES '<GIVEN ATTRIBUTE>' BELOW.

4f. Please indicate how important each of these aspects of a casual dining restaurant is to you. Please use a scale where <GIVEN ATTRIBUTE> has been given a score of 50 to show how important it is. Use this rating to judge all the other aspects. For example, if an aspect is only half as important as <GIVEN ATTRIBUTE> you might give it a 25, because that is half of 50. If an attribute is much more important than <GIVEN ATTRIBUTE> then you would give it a much larger rating. For example, if you think an aspect is three times as important as <GIVEN ATTRIBUTE> then you would give it a 150, and so on. There is no upper limit – use any number so long as it shows how important you think each aspect is. If you think something is not important at all, give it a 0. SHOW BELOW FORMAT WITH '50' FILLED IN FOR WHATEVER THE <GIVEN ATTRIBUTE> IS. IF THIS IS NOT POSSIBLE, PLEASE LET ME KNOW. RANDOMISE ORDER. ATTRIBUTE RATINGS CANNOT GO BELOW 0.

Prompt greeting	_____
Overall cleanliness	_____
Comfortable environment	_____
Server attentiveness	_____
Server friendliness	_____
Pace of meal	_____
Taste of food	_____
Temperature of food	_____
Receive bill in timely manner	_____
Reasonable prices	_____

Derived importance

FOR THE FINAL 2 QUESTIONS, CHOOSE ONE OF THE RESTAURANTS THAT APPEAR IN QUESTION 3. RESTAURANT CHOSEN BECOMES <RESTAURANT> IN QUESTIONS 5 AND 6.

5. Overall, how satisfied are you with your experiences dining at <RESTAURANT>? Compared to other casual dining restaurants, would you say you are …

[ ] Much more satisfied
[ ] A little more satisfied
[ ] About as satisfied
[ ] A little less satisfied
[ ] Much less satisfied

6. Please indicate how much you agree or disagree with each of the following statements about your dining experiences at <RESTAURANT>. RANDOMISE ORDER.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
I am greeted promptly at <RESTAURANT>	[ ]	[ ]	[ ]	[ ]	[ ]
Overall, <RESTAURANT> is clean	[ ]	[ ]	[ ]	[ ]	[ ]
<RESTAURANT> has a comfortable environment	[ ]	[ ]	[ ]	[ ]	[ ]
Servers at <RESTAURANT> are attentive	[ ]	[ ]	[ ]	[ ]	[ ]
Servers at <RESTAURANT> are friendly	[ ]	[ ]	[ ]	[ ]	[ ]
Meals at <RESTAURANT> are paced appropriately	[ ]	[ ]	[ ]	[ ]	[ ]
<RESTAURANT> has excellent-tasting food	[ ]	[ ]	[ ]	[ ]	[ ]
Food at <RESTAURANT> is served at the right temperature	[ ]	[ ]	[ ]	[ ]	[ ]
At <RESTAURANT> I receive my bill in a timely manner	[ ]	[ ]	[ ]	[ ]	[ ]
<RESTAURANT> is reasonably priced	[ ]	[ ]	[ ]	[ ]	[ ]

References

Allen, D.R. & Rao, T.R. (2000) *Analysis of Customer Satisfaction Data*. Milwaukee: ASQ Quality Press.

Bacon, D.R. (2003) A comparison of approaches to importance-performance analysis. *International Journal of Market Research*, **45**, pp. 55–71.

Bass, F.M. & Wilkie, W.L. (1973) A comparative analysis of attitudinal predictions of brand preference. *Journal of Marketing Research*, **10**, pp. 262–269.

Beckwith, N.E. & Lehmann, D. (1973) The importance of differential weights in multi-attribute models of consumer attitude. *Journal of Marketing Research*, **10**, pp. 141–145.

Bradley, R.A. & Terry, M.E. (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, **39**, pp. 324–345.

Cohen, S. (2003) Maximum difference scaling: improved measures of importance and preference for segmentation. *Proceedings of the 2003 Sawtooth Software Conference Proceedings*, San Diego, CA, pp. 61–74.

David, H.A. (1988) *The Method of Paired Comparisons* (2nd edn, revised). London: Charles Griffin.

Finn, A. & Louviere, J.J. (1992) Determining the appropriate response to evidence of public concern: the case of food safety. *Journal of Public Policy and Marketing*, **11**, pp. 19–25.

Fishbein, M. (1967) A behavior theory approach to the relations between beliefs about an object and the attitude toward the object. In: M. Fishbein (ed.) *Readings in Attitude Theory and Measurement*. New York: Wiley, pp. 389–399.

Gensch, D.H. & Recker, W.W. (1979) The multinomial, multiattribute logit choice model. *Journal of Marketing Research*, **16**, pp. 124–132.

Heeler, R.M., Okechuku, C. & Reid, S. (1979) Attribute importance: contrasting measurements. *Journal of Marketing Research*, **16**, pp. 60–63.

Holbrook, M.B. (1977) Comparing multiattribute models by optimal scaling. *Journal of Consumer Research*, **4**, pp. 165–171.

Jaccard, J.D., Brinberg, D. & Ackerman, L.J. (1986) Assessing attribute importance: a comparison of six methods. *Journal of Consumer Research*, **12**, pp. 463–468.

Kruskal, W. (1987) Relative importance by averaging over orderings. *American Statistician*, **41**, pp. 6–10.

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.

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

McFadden, D. (1973) Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (ed.) *Frontiers in Econometrics*. New York: Academic Press, pp. 105–142.

Myers, J.H. (1996) *Segmentation and Positioning for Strategic Marketing Decisions*. Chicago: American Marketing Association.

Myers, J.H. (1999) *Measuring Customer Satisfaction: Hot Buttons and Other Measurement Issues*. Chicago: American Marketing Association.

Orme, B. (2003) Scaling multiple items: monadic ratings vs paired comparisons. *Sawtooth Software Conference Proceedings*. San Diego, CA, pp. 43–59.

Retzer, J. (2006) The century of Bayes, *International Journal of Market Research*, **48**, 1, pp. 49–60.

Rosenberg, M.J. (1956) Cognitive structure and attitudinal affect. *Journal of Abnormal and Social Psychology*, **53**, pp. 367–372.

Soofi, E.S., Retzer, J.J. & Yasai-Ardekani, M. (2000) A framework for measuring the importance of variables with applications to management research and decision models. *Decision Sciences Journal*, **31**, pp. 596–625.

Stapleton, L.M. & Edmonds, M. (2005) An exploration of the validity of the unbounded write-in scale for inter-individual research. *International Journal of Public Opinion Research*, **17**, pp. 484–494.

Stephenson, W. (1953) *The Study of Behavior: The Q-Technique and its Methodology*. Chicago: University of Chicago Press.

Strathmann, W.C., Zastrowny, T.R., Bayer, L.R., Adams, E.H., Black, G.S. & Fry, P.A. (1994) Patient satisfaction surveys and multicollinearity. *Quality Management in Health Care*, **2**, pp. 1–12.

Swait, J. & Louviere, J. (1993) The role of the scale parameter in the estimation and use of multinomial logit models. *Journal of Marketing Research*, **30**, pp. 305–314.

Theil, H. & Chung, C. (1988) Information-theoretic measures of fit for univariate and multivariate linear regressions. *American Statistician*, **42**, pp. 249–252.
Thurstone, L.L. (1927) A law of comparative judgment. *Psychology Review*, **34**, pp. 273–286.
Wilkie, W.L. & Pessemier, E.A. (1973) Issues in marketing's use of multi-attribute attitude models. *Journal of Marketing Research*, **10**, pp. 428–441.

About the authors

Keith Chrzan is the Vice President of Marketing Sciences at Maritz Research. Previously he was Director of Marketing Sciences at IntelliQuest, Inc. He has also spent time on the client side with Boehringer Mannheim Diagnostics and Bayer, Inc. He speaks frequently at industry conferences, usually on topics surrounding the design and analysis of choice experiments. He has a Bachelors degree in Philosophy of Religion from the University of Notre Dame and an MBA in Marketing from Indiana University.

Natalia Golovashkina is a Senior Research Analyst at Maritz Research. Previously she was a visiting assistant professor in the Freeman School of Business at Tulane University. She completed her PhD in Management at Cornell University in January 2005. Her dissertation focused on evaluation of performance of contracts in supply chains.

Address correspondence to: Keith Chrzan, 996N 250E, Chesterton, IN 46304, USA.

Email: keith.chrzan@maritz.com