validity coefficients than models that also incorporated measures of expectations.

The 0-10 point “describes/does not describe” scale, however, is not effective when measuring ideal point attributes. One method that can be used to measure an ideal point attribute is to have respondents evaluate the performance of a service (e.g., on a unit of measurement that reflects different “amounts” of the service). This type of scale is called an intensity scale.

For example, in measuring the ideal frequency with which a corporate account calling officer should make a personal visit to a customer’s place of business, one could ask respondents the ideal frequency with which these visits should occur. Then, respondents could be asked the actual frequency with which these visits do occur. One could then subtract the actual from the ideal frequency to examine the “gap” between actual and ideal performance.

An alternative approach is to use a scale that measures evaluated performance. For example, if a performance attribute is described on an evaluative continuum such as poor performance vs. excellent performance, the assumption can be made that infinite ideal points are involved. That is, excellent performance can be assumed to be preferred over poor performance.

This is in contrast to intensity-type scales in which the amount of the attribute is measured. Once performance is measured on an evaluative scale, the ideal point is assumed to be infinite, eliminating the problem and therefore the need to measure ideal points.

Perceived performance scales, evaluated performance scales, or intensity scales might provide better measures of quality than do expectations or requirements scales. However, the process of testing the reliability and validity of scales discussed by Devlin and colleagues is an important contribution to the literature and can be used in future theoretical and applied research to discover how scales in general should be constructed and administered to develop the most valid measures of quality.

Terry Grapentine is president of Grapentine Co., a marketing research firm based in Ankeny, Iowa.

---

The Measurement Imperative

Empirical results on the number of scale points help assess customer satisfaction.

By Dick R. Wittink and Leonard R. Bayer

Susan Devlin, H.K. Dong, and Marbue Brown provided a useful discussion of considerations relevant to the selection of a scale for measuring quality in their Summer 1993 Marketing Research article. Although they made reference to a few statistical results in their comparison of alternative scales, their arguments were primarily conceptual. In addition, they did not discuss the influence of measurement scales on the reliability or validity of the estimated impact of controllable variables.

Managers need access to customers’ perceptions of how firms perform on controllable variables as well as how these variables relate to (or influence) overall satisfaction. Knowledge about the impact of controllable factors guides managers to focus their attention on making the right improvements. Only with such guidance can they achieve continuous and predictable improvements in customer satisfaction.

For one firm, customer satisfaction has been a management objective for several years. During this time, the firm has experimented with two alternative measurement systems, which we refer to as A and B.

- System A consists of a series of 5-point scales on which customers report their degree of satisfaction with individual components (controllable variables) and their degree of overall satisfaction.
- System B employs dichotomous (2-point) scales on which customers report whether they have experienced problems (yes or no) with individual components. Overall satisfaction in system B is measured with a 10-point scale.

Our approach and the results we obtained were adopted by the firm for which the study was done, and the system we recommended—system B—was put in place.

In the criteria we adopted for comparing alternative mea-
In comparing two customer satisfaction measurement systems, the authors favor a system with a 10-point scale for overall satisfaction and 2-point scales for individual items over a system with 5-point scales for everything. Although results may not generalize to all contexts, the approach can be used by researchers to make similar comparisons for firms or clients.

Measurement scales for “scoring” overall customer satisfaction, the score obtained in a particular survey of customers is the number used for rewarding managers. For each criterion, we show how alternative measurement systems can be compared and reveal the results we obtained for this particular firm.

Although we find that system B is consistently favored over system A on all the criteria, we do not claim that such a result will occur for all firms, industries, instruments, methods of data collection, etc. The criteria and the procedures we have adopted for the comparison, however, should be useful for similar comparisons in other contexts. We examine reliability, statistical power, and measurement sensitivity.

Reliability. Holding other things constant, the required sample size is especially dependent on the standard deviation of the overall satisfaction scores across respondents.

To maintain comparability across scales that differ in the number of scale values, we specify that the tolerable difference (D) for the 10-point scale has to be (10-1/5-1) or 2.25 times the tolerable difference for the 5-point scale. Thus, we use the maximum possible difference for each scale to obtain a ratio for converting the tolerable difference for one scale to the tolerable difference for the other scale.

It follows that if the standard deviation for the 10-point scale is less than 2.25 times the standard deviation for the 5-point scale, the required sample size for the 10-point scale is less than what is required for a proportionally equivalent amount of precision for the 5-point scale.

For the firm in question, 1.9 is the typical standard deviation for the 10-point, and 1.0 for the 5-point scale (based on sample sizes in excess of 100,000). Thus, the 10-point scale is favored on the sample size criterion for reliable measurement of average overall satisfaction. With the observed standard deviations, the 10-point scale requires only 71.3% of the sample size required for the 5-point scale.

We show a few arbitrary sample size calculations in Exhibit 1 to illustrate the difference in requirements between the two scales, based on the empirical standard deviations of 1.0 and 1.9. If the sample sizes were the same for the two measurement scales, the 10-point scale would afford a higher degree of precision on a relative basis than would the 5-point scale.

Power. The measurement system adopted also should provide the maximum opportunity to detect changes in average overall satisfaction in the population of all customers if changes occur.

For example, if, in 1994, the true but unknown average satisfaction is higher than it was in 1993, the measurement system that has the highest chance of detecting the improvement for a given sample size is the preferred system. By making some simplifying assumptions (e.g., independent samples, constant sample size over time, no change in the standard deviation over time), it is possible to show that the 5-point scale has 71.3% of the power of the 10-point scale.

Thus, just as with the reliability comparison, if the two measurement scales have the same sample sizes, the 10-point scale provides a greater opportunity to detect changes in overall satisfaction, given the observed standard deviations.

Sensitivity. The sensitivity of each scale to changes in individual components that can be assumed to be related to overall satisfaction (e.g., product quality, service quality) is indicated by the opportunity for increases in overall satisfaction to occur. This is defined as the percent of overall satisfaction scores below the maximum possible.

Based on the same data referred to previously, about 44% of the respondents providing answers on the 5-point scale scored the company below 5. Similarly, on the 10-point scale, based on data collected in the same year, about 82% of the respondents scored the firm below 10. Thus, the proportion of customers for whom overall satisfaction can be improved is much greater for the 10-point than for the 5-point scale. (Extremely large sample sizes guarantee that these percentages differ significantly from each other.)

One may argue that, by definition, the best score on a 10-point scale is not comparable to the best score on a 5-point scale. For example, it is quite likely that most respondents scoring the firm 5 on the 5-point scale score it either 9 or 10 on the 10-point scale. It would indeed be very surprising if all respondents scoring the firm 5 gave it a 10 on the wider scale.

But this is precisely the advantage of a wider scale. Unless one can argue that the difference between scores 9 and 10 is irrelevant, the scale should be as wide as possible, subject to

<table>
<thead>
<tr>
<th>Confidence</th>
<th>5 Points</th>
<th>10 Points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D</td>
<td>n</td>
</tr>
<tr>
<td>95%</td>
<td>0.1</td>
<td>384</td>
</tr>
<tr>
<td>99%</td>
<td>0.1</td>
<td>663</td>
</tr>
<tr>
<td>95%</td>
<td>0.05</td>
<td>1,537</td>
</tr>
<tr>
<td>99%</td>
<td>0.05</td>
<td>2,652</td>
</tr>
</tbody>
</table>

* The standard deviations are assumed to be 1.0 for the 5-point scale and 1.9 for the 10-point scale.
the ability of respondents to deal realistically with the range of values provided. With the 10-point scale, managers will be more motivated to identify additional opportunities to increase customers’ overall satisfaction scores.

To provide a more equitable comparison on the sensitivity criterion, we use the concept of relative sensitivity. On the 5-point scale, average satisfaction was 4.334, which leaves an average improvement opportunity of .666 (5-4.334) compared to the theoretically maximum possible of (5-1) = 4. Thus, on the 5-point scale, .666/4 = 16.7% of the maximum possible improvement is still to be accomplished.

On the 10-point scale, average satisfaction was 7.983, which leaves on average 2.017 out of the maximum of (10-1) = 9. This gives a relative sensitivity measure of 2.017/9 = 22.4%. Therefore, even on this relative basis, there is more room for improvement with the 10-point scale.

The occurrence of strongly skewed distributions of overall satisfaction scores has been identified as a problem in customer satisfaction research. In a 1992 Journal of Marketing article, Claes Fornell mentioned that, in Sweden, where a national customer satisfaction barometer index to complement productivity measures is in use, 10-point scales were adopted, in part, to minimize skewness in the distribution of satisfaction scores.

For the data used here, we find that on a relative basis, the average overall satisfaction score for the 10-point scale is farther from the maximum than the average score for the 5-point scale (leaving more room for improvement). At the same time, the standard deviation satisfaction score is smaller for the 10-point than for the 5-point scale on this relative basis (which enhances reliability and power). The relative values for these two measures together substantiate the claim that skewness in the data is a more severe problem for the satisfaction scores on the 5-point than on the 10-point scale.

Finally, we must note that we assumed the scores for both measurement scales to be metric. This assumption was necessary to justify the calculation of (arithmetic) averages and standard deviations.

**Impact of Controllable Variables**

The potential gain from improvement in a controllable variable on overall satisfaction is a function of two quantities: (1) the average customer perception of performance on the variable and (2) the influence of improvement in the variable on overall satisfaction for the customer.

Performance is measured directly by asking customers to indicate on a scale their experience with the variable(s) in question. Influence or importance, however, is more difficult to determine. Self-explicated importance questions suffer from several limitations. For example, respondents’ interpretations of importance might be unclear unless reference is made to a precisely defined difference in possible performances on a variable.

Even if self-explicated weights are obtained for well-defined differences in individual attributes, there is no obvious way to achieve scale compatibility in the importances and the performance measures of the variables. As a result, in many customer satisfaction measurement systems, the importances of the variables with regard to their influences on overall satisfaction are inferred via the slope coefficients from multiple regression analysis.

Finally, because the individual variables typically are correlated, it’s inappropriate to use a simple regression slope coefficient. (In some cases, simple correlation coefficients—between overall satisfaction and each of the individual components—are used. However, it’s easy to show that the magnitudes of correlations have no managerial relevance.)

To determine the validity of a multiple regression, some researchers use the $R^2$ value. For example, one might prefer the measurement system that generates the highest explanatory power. In our analysis, the $R^2$ value is greater for the equation with 5-point scale values than for the 10-point scale values. However, due to the difference in the criterion variables (system A having a 5-point scale, system B having a 10-point scale), the $R^2$ values are not comparable. Even if the comparison could be made, the total explanatory power of the equation is not at issue; what matters is the ability to estimate valid (and reliable) separate effects of individual components.

To compare systems A and B, we propose the following three criteria for multiple regression results:

**Validity.** We first determined for both measurement systems that the multiple regression model is significantly better than an equation with equal weights for the individual components. We also found for both systems that the expected predictive validity for each multiple regression equation is superior to the expected predictive validity for an equal weight model. Thus, we can proceed to interpret the individual slope coefficients.

To establish the validity of the coefficients in a multiple regression analysis, we need to know the true parameter values. Obviously, the parameters are, in general, unknown. However, in our case all components are defined in such a way that the true effect of each component, holding the other ones constant, is nonnegative. Therefore, we define face validity in terms of the number of nonnegative slope coefficients.

A multiple regression analysis involving 11 components, comparably defined for systems A and B—applied separately to a sample of 1,000 observations for each system—provided the following results: For system A (5-point scales), 7 of the 11 coefficients were nonnegative, whereas, for system B, all 11 were nonnegative. Thus, this simple face validity comparison favors system B.
One might wonder if the difference in face validity between systems A and B is due to respondents approaching the task differently between the two systems or to an artificial difference in the number of scale values. To consider this, we recoded the 5-point scale predictor variables in system A to dichotomous scales. (The recoding algorithm was chosen such that the frequency distribution for the recoded scale from system A and the observed frequency distribution for system B produced the minimum $\chi^2 >$ value.) Based on the recoded data, we obtained 8 out of 11 nonnegative coefficients.

One also could ask to what extent the 10-point overall satisfaction scale artificially increases the validity of the results relative to a 5-point scale criterion variable. To examine this, we collapsed the 10-point scale in system B to a 5-point scale, and found that the validity of the slope coefficients for system B was not affected; all 11 coefficients were still nonnegative.

Therefore, we conclude that the greater face validity for system B was not due to a scale artifact but rather to a systematic difference in the way respondents approached the task. For example, more effort is required from a respondent who has to score a firm on each of many components on a 5-point scale. Because of respondents’ desires for minimum effort, they may tend to rate the components similarly. The greater the similarity of scores across the components for each customer, the higher the multicollinearity and the more difficult it becomes to obtain the expected signs for all slope coefficients. Thus, system B outperforms A in terms of face validity.

**Reliability.** The absolute reliability of the individual coefficients is indicated by the estimated standard errors of the slope coefficients in each multiple regression analysis. However, because the units of the criterion and predictor variables are not comparable, relative reliability measures are needed for comparison purposes.

We used the ratio of the standard error of the coefficient to the corresponding slope coefficient (equivalent to the coefficient of variation). This ratio is unitless and, hence, comparable across systems. The smaller this ratio, the higher the (relative) reliability.

For system A, this ratio varies from -3.5 to 4. The average of the positive values equals 1.1, whereas the average of all values (using absolute values for the negative ones) is 1.7. For system B, the ratio ranges from .1 to 5. The average of all 11 values equals .8. Thus, the average of these reliability measures, whether one ignores the negative coefficients occurring in system A or not, favors system B.

This means that, for the model with 11 components considered here, and equal sample sizes, the power (opportunity to detect an effect if one exists) is higher for system B than for system A. These average relative reliability measures change only slightly if the values are recoded to equate the number of scale values for systems A and B. Thus, this difference in relative reliability also is not due to an artificial difference in the number of scale values.

We also determined the stability of the coefficients when one predictor is omitted. We conducted this analysis for four predictors, omitted one at a time, and computed the change in the value of each coefficient relative to its value in the analysis with all 11 predictor variables.

A convenient way to summarize instability is to count the total number of times this relative change is 100% or more. For system A, this is 14 times (for the recoded data, it is 13 times); for system B, it is only 3 times. Thus, on this measure of stability the conclusion is the same: System B provides better data.

Note that the relative reliability of the estimated effects is determined by various factors, including the explained variance of the equation and (lack of) multicollinearity among the predictor variables. The covariation between the criterion variable and a linear combination of the predictor variables may be partitioned into a component that is unique to the specific predictor variables and a component that is shared among the predictor variables. Multicollinearity causes this shared variance component to be very large. This shared variance is responsible for the variance inflation factor that influences the estimated standard errors of the coefficients. The higher multicollinearity for system A results in a higher coefficient of variation for the results from system A.

**Managerial relevance.** The interpretation of the multiple regression result for dichotomous predictor variables ($0 = \text{problem}; 1 = \text{no problem}$) is that a given slope coefficient indicates the predicted change in overall satisfaction if a given customer is changed from experiencing a problem to experiencing no problem, holding the other predictors constant.

To determine the maximum possible impact on overall satisfaction that’s achievable if all respondents who indicate a problem are moved to indicating no problem, we multiply the slope coefficient by the proportion of respondents currently indicating a problem for each predictor variable.

This measure appropriately combines the two aspects that indicate the effect ($b_1$) and the opportunity ($P_1$) for each component. Assuming equal cost for achieving “no problem” on any component, this measure can be used to identify the “most promising” component(s). Thus, this measure is very actionable.

Finally, the cost of converting respondents from problem to no problem or improving the performance is different for each component, and this factor can be taken into account, for example, by dividing the impact by a cost estimate. The resulting measure could provide an impact per dollar for each component.

We have established that, for one firm, a customer satisfaction measurement system based on a 10-point overall satisfaction scale and 2-point scales for individual components is superior to a system that uses 5-point scales for overall satisfaction and individual components. For scoring, we found that the 10-point scale requires a smaller sample size for a given amount of reliability and power and leaves more room for improvement.

For determining the impact of controllable variables, we obtained higher face validity and reliability in the results for the system using a 10-point overall satisfaction scale and 2-point scales for the components than for the system with 5-point scales.

We also determined that the advantage for the former system is not due to artificial differences created by the number of scale values. Rather, it appears that if a 5-point scale is used for all components, respondents are inclined to create similiar-
ties in their responses across the components. This similarity contributes to multicollinearity.

The results we observed for this one firm will not necessarily translate to other firms. We strongly encourage other researchers to make similar comparisons and report their findings.

We also recommend that users of customer satisfaction research insist on establishing the predictive validity of results. Managers should determine that the value of a satisfaction measurement system is greater than its cost. One way to establish this is to provide diagnostic information to a randomly chosen subset of managers. Useful questions to be addressed are (1) Do the decisions made by managers who have access to customer satisfaction information differ from those of managers who do not have access to such information? and (2) Are subsequent results better for the managers with access than for those without access?

There are important reasons why the results from such an exercise are not straightforward. For example, customers may differ substantially in the sensitivity of overall satisfaction to improvements in individual components. The types of satisfaction measurement systems we have briefly reviewed here can accommodate such customer heterogeneity only to a limited extent. It is well-known that pooling data across respondents who are heterogeneous in their sensitivities to improvements in individual components is likely to result in biased estimates.

Dick R. Wittink is professor of management and marketing at Yale University, school of management. Leonard R. Bayer is executive vice president, chief scientist, and member of the management committee and board of directors at Harris Interactive.

## The “Perfect” Scale

The number of points on a scale is less important than its application.

*By Diane H. Schmalensee*

With so many businesses recognizing the importance of customer satisfaction, many researchers are taking a serious look at the viability of their own measurement methods. The articles under discussion in this issue illustrate the industry’s struggle toward refinement. Given the variety of scales tested and the criteria used by the authors, it’s not surprising that they all reach different conclusions.

After reading these articles, I have drawn several general conclusions:

- It’s virtually impossible for a single article to consider all the possible combinations of labels and number of scale points. There are simply too many possible combinations to be analyzed at one time.
- The choice of scales is highly dependent on the criteria used. Although all the authors’ criteria are important to them, the criteria themselves are highly subjective.
- The choice of scales may be (and should be) strongly influenced by the research objectives when measuring customer satisfaction.

Some case examples may help illustrate why the choice of scales is so dependent on research objectives and market conditions. Rather than using the 4-, 5-, and 10-point scales under contention, I’ll use a neutral example of a 7-point scale in most cases.

### Fewer Strong Competitors

Company 1’s customer satisfaction goal was to outperform its two best and largest competitors to build customer retention and accelerate customer acquisition. Rather than use a comparative scale label, such as “performs better or worse than competitor 2 or 3,” this company chose to use a simple 7-point performance label (from “poor” to “excellent”), with respondents rating each company separately on each attribute.

This led to a performance comparison table that highlighted competitive strengths and weaknesses. Executives were able to see quickly that Company 1 was superior on variables A and B, in a dead heat on vari-