Selecting a Scale for Measuring Quality

The perfect rating scale doesn't exist, but some produce more reliable and valid findings than others.

By Susan J. Devlin, H.K. Dong, and Marbue Brown

In the quest to measure the quality of products and services from the customers' perspective, researchers should assess the quality of the measurement tool as well. One focus for such an assessment is the response scale that customers would use to communicate their views about the performance of the product or service. What is the best scale? Which categories are positive indicators of performance? Can responses be interpreted to drive quality improvement? The authors address these questions, discuss the pros and cons of various scales, and provide guidelines for assessing rating scales.

One approach to measuring the quality of products and services is to ask customers to comment on their experiences and analyze and interpret verbal responses (i.e., open-ended verbatims). An alternative is to ask for assessments of different aspects of or events in the product or service delivery experience using a rating scale.

Although both approaches play a role in assessing quality and are optimally used together, the use of a rating scale allows researchers to calibrate responses across customers and interpret and present results efficiently using sophisticated statistical tools.

Since 1987, Bellcore's Measurements Research Group, on behalf of telecommunications companies in the U.S., has tested a variety of rating scales in measuring diverse products and services such as telecommunications services (e.g., installation, repair, billing, marketing, and directory assistance), software systems, research and consulting services, and internal corporate support functions (e.g., secretarial services and conference planning).

Leading companies in many industries are seeking measurements to help them identify their strengths and weaknesses to become more competitive and profitable. A necessary first step in this endeavor is to ensure the quality of the measurement tool. The choice of a rating scale not only affects the reliability and validity of the findings, but also influences how results are used and how easy the survey is to answer and administer.

Too frequently, however, a scale is "borrowed" or made up without regard to its effectiveness as a measurement tool, resulting in biased or misleading results. There seems to be as many scales as there are companies conducting surveys.

Although selecting an appropriate rating scale is a necessary first step in developing a good measurement instrument, establishing statistical reliability and validity through a multistep testing and retesting process should be accorded the highest priority in selecting a scale.

What Makes A Good Scale?

A good rating scale should have the following characteristics:

- Minimal response bias.

Exhibit 1

Commonly used rating scales

Satisfaction scale

2-Satisfaction	4-Satisfaction	5-Satisfaction	5n-Satisfaction
1=Satisfied	1=Very satisfied	1=Very satisfied	1=Very satisfied
2=Dissatisfied	2=Satisfied	2=Satisfied	2=Satisfied
	3=Somewhat dissatisfied	3=Somewhat satisfied	3=Neither
	4=Very dissatisfied	4=Dissatisfied	4=Dissatisfied
		5=Very dissatisfied	5=Very dissatisfied

Performance scale

4-Excellence	5-Excellence
1=Excellent	1=Excellent
2=Good	2=Good
3=Fair	3=Just OK
4=Poor	4=Poor
	5=Terrible

Gap scale

5-Expectations	4-Requirements
1=Much better	1=Exceeded
2=Better	2=Met
3=Just as	3=Nearly met
4=Worse	4=Missed
5=Much worse	

Non-anchored scale

Grade	Number
A	10
B	9
C	-
D	1
F	0

- Respondent interpretation and understanding.
- Discriminating power.
- Ease of administration.
- Ease of use by respondents.
- Credibility and usefulness of results.

The sample scales in Exhibit 1 define and exemplify these criteria.

Minimal Response Bias

The way respondents use scales can negatively affect our ability to interpret results. Two well-known effects are positive response bias and endpoint avoidance.

Positive response bias: Most people want to be "nice," tending to select the less-critical judgmental position. Consider the 4-satisfaction scale, many "satisfied" customers subsequently give negative comments in their verbatims, leading service providers to misinterpret their competitive position. Service providers may be at risk of losing these mildly disappointed customers as they are less likely to take the initiative to voice concern and may quietly seek another supplier.

Adding a neutral or "politely negative" category (e.g., the 5-satisfaction and 5n-satisfaction scales) shifts response away from "satisfied" with little affect on the two "dissatisfied" categories. For example, in a parallel trial of the 4-satisfaction and 5-satisfaction scales in assessing several support services, John Hughes of Bellcore found the percentage of positive response ("very satisfied" plus "satisfied") dropped 0% to 8%, with variation dependent on the size of the "satisfied" category in the 4-satisfaction scale.

Some researchers argue against including a midpoint or neutral category, forcing a respondent to take sides. Support is strong for this approach with general preference research where there is no positive or negative connotation. However, with judgmental or performance assessments, forcing sides usually results in a positive response, which leads service providers to misinterpret results. Using a neutral or midpoint rating doesn't necessarily result in an odd number of points in a scale. For example, the "nearly met" category in the 4-requirements scale serves exceptionally well.

Endpoint bias: Some respondents avoid endpoints in a scale. This phenomenon further exaggerates positive response bias and discourages the use of a two-point scale. More positive responses result with the 2-satisfaction vs. the 4-satisfaction scale, for example. Although a two-point scale or a check-off is sometimes effective when simplicity is required, it must be very carefully tested.

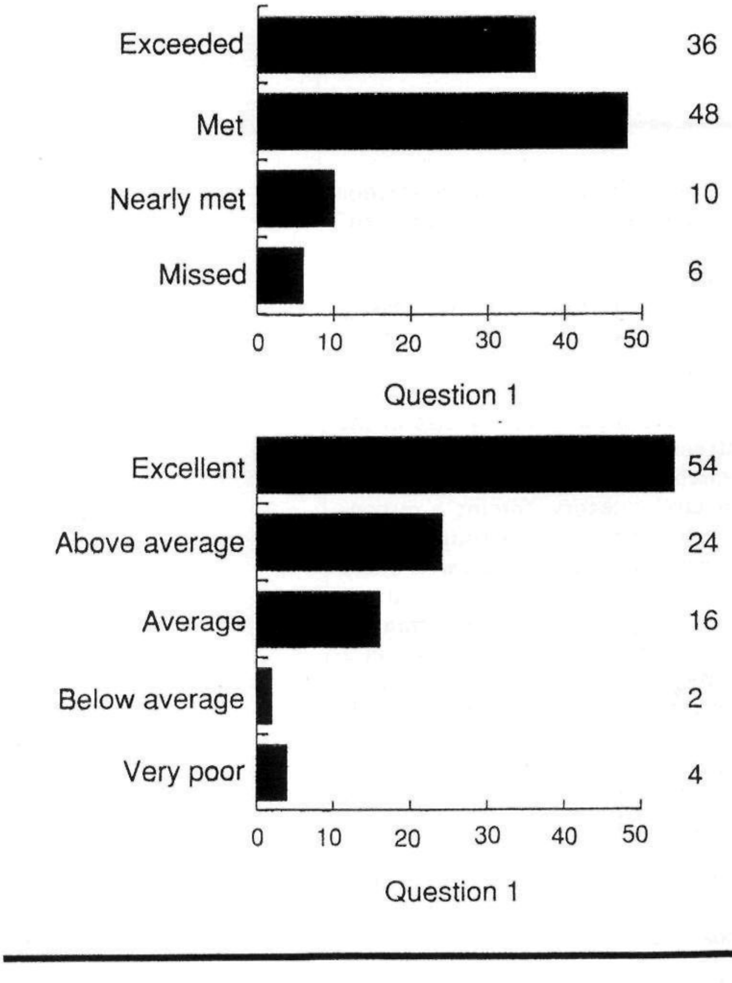
Respondent Interpretation and Understanding

Respondents' interpretation of the scales' categories should not be assumed; colloquial use, as well as dictionary meanings, must be taken into account. For example, consider one of the most commonly used scales, 4-excellence. As Stanley Payne noted as early as 1951, "fair" has conflicting colloquial uses. "Fair" weather implies a positive meaning, and a "fair" ruling implies some form of neutrality. Interpretation is further confounded by geographic differences; for example, in some pockets of the South, "fair" is better than "good," whereas it tends to have a mediocre connotation in the Northeast and elsewhere.

Consider another example. Some researchers, who have examined verbal comments made by respondents in conjunction with a scaled response, conclude that "satisfied" is an inferior criterion to "excellent," perhaps implying that a supplier is meeting commitments rather than customers' expectations.

Exhibit 2

Examples of how respondents use different scales



Some researchers discard verbally anchored scales in favor of numerical or lettered scales (e.g., the grade and number scales in Exhibit 1). Unfortunately, they face similar problems in interpretation. Which categories are positive? An adequate grade to one person might be a B or a 70% (7), but to another it is a C or 50% (5). Further, some people are not numerically or spatially oriented; for them, translating thoughts and feelings to a number line is a difficult and perhaps unreliable task. Similar arguments can be made against purely verbal scales for those who are more numerically oriented.

Discriminating Power

The ability to discriminate between degrees of customer opinion can help distinguish which service levels are poor vs. adequate vs. exceptional/desirable. Subsequently, this may lead to differentiating a competitive vendor from favored-partnership status. The latter means the supplier is protected from a competitive threat. If diverse service experiences translate to one response category, information is lost.

More scale categories are not necessarily better, but at least three points should be used. Research by Jacob Cohen, for example, has shown that the use of a two-point scale (i.e., yes/no or satisfied/dissatisfied) can lead to 20%-66% loss in prediction (e.g., R^2) and statistical power. At the same time, other researchers have suggested that little is added beyond five points.

For example, a summary of responses to questions with a 10-point response scale may show all categories used, but this does not mean that a refined gradient in performance has been assessed. If each respondent uses no more than five scale categories, this suggests that respondents—when faced with excessive number of categories—are defining a subset of the scale. Rather than discriminating different levels of performance, a calibration problem is introduced. Another source of response error is introduced when using verbal scales on the telephone: People can't remember the choices if there are too many of them.

The ability to measure is affected not only by the number of categories, but also by the nature of the scale. The goal is to have well-spaced anchor points (interval scale) representing the possible range of opinions about the service.

Exhibit 2 summarizes the results from parallel trials of a telecommunication provisioning survey using different scales. These results indicate that the 4-requirements scale is more discriminating than the 5-excellence scale. Similar results were found by Hughes and when we tested the 5-expectations scale against satisfaction and excellence scales.

Even when the percentage of positive response for a satisfaction or an excellence scale matches that for a requirements or expectations scale, the percentage of response for the top category, "very satisfied," is greater, thus indicating it is less discriminating.

Even more telling, statistical estimates of the validity and reliability of questions tend to be strong, dimensions of quality can be identified easily, and correlations between questions (multicolinearity) are controllable with the 5-expectations and the 4-requirements scales. Typically, the reliabilities (*coefficient alpha*) of questionnaires are in the .80 to .90 range and the validities (*multiple R*) in the .60 to .80 range, both of which are above industry standards.

Ease of Administration

The same scale may not work equally well on paper and in a telephone interview. For example, the midpoint of the 5n-satisfaction scale, "neither satisfied nor dissatisfied," is effective on paper, but respondents avoid it in a telephone interview, thus distorting the meaning of surrounding categories.

The order in which a scale is presented can help counterbalance positive response bias. As respondents tend to check off the first category that seems relevant on a paper survey, placing negative categories first (to the left) helps cancel out the tendency to respond positively. Since respondents on the telephone may favor the later categories heard, similar reasoning suggests reading a scale from positive to negative. Thus, a scale's order, position, and presentation all can affect its delivery.

Ease of Use by Respondents

It's important for respondents to use the scale accurately, without problems and irritation. This does not mean that they have to "like" the scale. For example, one method of evaluating scales that we found of little value is to have a group of respondents discuss what they think of the possible scales. Contemplating a scale is not the same as their quick reaction to a survey. Furthermore, recommendations—"Just ask me yes or no"—do not reflect actual unconscious use.

Several checks identify when respondents may be having a problem with or are irritated by a scale:

- Does the interviewer have to repeat a scale frequently? Are many respondents using phrases that are not categories in the scale, or are they asking what the scale means?
- Is one category of the scale rarely used, compared to those on either side?
- Are the survey dropout rates or individual question non-response rates unusually high?
- Do too many customers use the same response category for all questions?
- Are many comments inconsistent with the responses?

Credibility and Usefulness of Results

The primary goal of measurement is to gather useful information to identify service improvement opportunities and motivate employees. If employees continually challenge the interpretation of a scale—even without justification—then energies are redirected away from quality improvement.

This does not mean that employees should dictate scale categories or that their interpretations should take precedence over those of customers in defining a scale's meaning. But researchers may need to reject or replace a scale if the obstacles are too great.

Review of Selected Scales

There is no perfect scale, only bad and better ones. Here are some comments on each type of scale presented in Exhibit 1:

2-satisfaction: Fails response bias and discriminating power criteria. However, combining a strong scale for overall assessments with check-off lists for identifying detailed problem areas has proven to be an effective compromise when trying to control the time commitment required for long surveys that may be needed to assess complex service offerings.

4-satisfaction or 4-excellence: Both fail because of positive response bias and weak discriminating power.

The choice of an appropriate rating scale for measuring product or service quality is a necessary first step in developing a quality measurement tool.

5-satisfaction or 5n-satisfaction: Both address positive response bias, but fare less well than excellence, expectations, and requirements scales in discriminating high-end performance. 5n-satisfaction also fails in telephone delivery.

5-excellence: Scores well in all criteria; however, it is suboptimal in discriminating power because of a tendency for response distribution to be more skewed to the positive end of the scale when compared to the expectations or requirements scales. This scale is favored when expectations are not likely to be well formed in customers' minds prior to the product/service experience. Robert Westbrook's 7-point "delighted" to "terrible" scale is another alternative for a paper survey, but not for a telephone survey.

Grade: Suffers from inconsistency in interpretation across respondents and company employees.

Number: This also suffers from inconsistency of interpretation. In addition, it has too many categories. A smaller number of categories with phrases to anchor mid- and endpoints may be useful on a paper survey to address both numerical/spatial and verbal orientations. However, this cannot be accomplished on the telephone, and the verbal anchors need to be semantically assessed as with the other verbal scales which we have described.

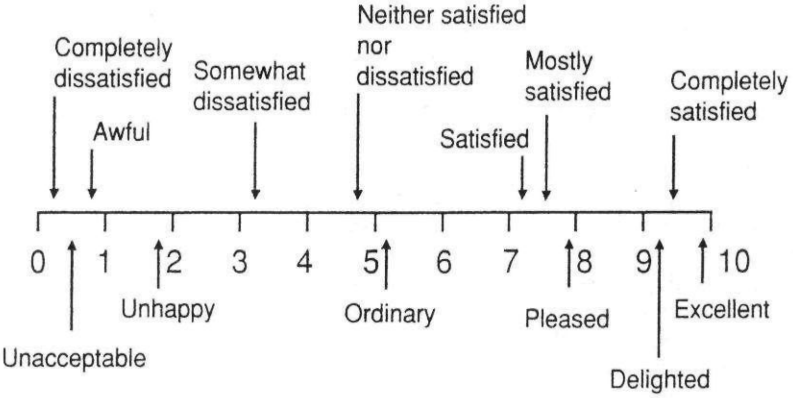
5-expectations and 4-requirements: These two scales are strong on all criteria. In both extensive research and diverse practical applications, they have led to the most detailed assessments. We assess these scales together because neither is clearly favored. A comparison of their tradeoffs is useful in demonstrating that there is no optimal or best scale.

Both scales counteract response bias well. "Just as expected" and "met" are nearly always clear, positive responses; however, the 4-requirements scale has a slight edge as the "nearly met" category serves as a "politely negative" response exceptionally well.

The two scales consistently result in high reliability and validity scores (*coefficient alpha* = .80-.90; *multiple R* = .60-.80) for questionnaires, discriminating service levels without being too difficult to use by respondents. The "much better" and "better" categories effectively isolate superior service; less is known yet about similar success with the "exceeded" category. Both have fairly high consistency across divergent customer groups in how the scale is interpreted.

Exhibit 3

Where do people place response categories on a scale of 0-10?



Care must be taken in wording some questions to ensure "exceeded" is not misinterpreted as "too high cost" or "too long." At the same time, the 5-expectations scale requires careful wording of the introduction to ensure expectation is associated with customers' needs.

Both scales work well in telephone and paper delivery. However, the 5-expectations scale may require more repeating in a telephone interview to be comfortably used but not to the point of being a deterrent. Both scales have been well-received in companies where they have been introduced because they link measures to current definitions and philosophy about total quality management.

Testing A Scale

Even though starting with an established scale is best, any scale should be tested or retested to ensure that it is reliable and valid. It also should work well with a particular set of questions, a particular group of customers, and a particular survey delivery method. If changes are made, retest to confirm performance. Parallel trials are a critical tool to assess changes and make other comparisons. Some of the analytical techniques that have proved fruitful in evaluating scales follow.

Semantic difference test: This test evaluates how people interpret and order phrases. Subjects are asked to place a response category or phrase on a numeric scale, e.g., 0-10. Exhibit 3 shows the average response to satisfaction categories of about 100 Bellcore employees to the placement of some of the 19 phrases studied. The dispersion of responses (not shown here) and the proximity of different words helps define a scale.

This study resulted in early use of the 4-satisfaction scale with support services measures. A semantic difference test helps ensure a wide range of opinions are covered, and they are well spaced.

Bar plots of response patterns for each question: This is a check to see if all categories are used, if unexpected patterns occur, or if there are too many nonresponses or "don't knows."

Bar plots of respondents' usage: Tabulating and plotting the number of categories respondents use and the percentage favoring each category can uncover such problems as bias towards or away from one category and frustration with a scale, which may be reflected in respondents thoughtlessly repeating one category to rush through the survey.

Some people legitimately use only one category across all questions. This occurrence increases as the number of questions decreases and the quality of the product or service increases. However, an unusually high percentage of patterned responses suggests potential problems with the scale or questionnaire.

Monitor telephone delivery: Monitoring interviews can uncover problems from a respondent's tone of voice, inquiries about the scale, hesitation, or quick, thoughtless responses.

Verbatim analysis: During trials, the interviewer can stimulate comments with a neutral "why" after a response, and researchers can code comments later for type of issue and tone of voice. (More directed requests for comments are introduced after testing to optimize the quality improvement opportunities detected.) The number of positive and negative comments and the nature of the comments suggest how a respondent interprets the category and whether there is consistency between the categorical response selected by a respondent, the intended meaning of the category, and respondents' verbatims.

Advanced statistical analyses: The use of multivariate statistical methods, such as factor analysis, logistic regression, and covariance structure analysis, helps assess which scale delivers the highest reliability and validity measures, reduces multicolinearity concerns, and has the greatest prediction power of criterion measures (e.g., overall quality or loyalty).

Conclusion

The choice of an appropriate rating scale for measuring product or service quality is a necessary first step in developing a quality measurement tool. Several factors must be given special attention in evaluating scales: response bias, ease of use and administration, and actionability of results. Above all, however, a good scale is one that is reliable and valid. Establishing reliability and validity through a multistep testing and retesting process should be afforded the highest priority in selecting a scale.

Additional Reading

Churchill Jr., Gilbert A. (1979), "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research*, 16 (February), 64-73.

Cohen, Jacob (1983), "The Cost of Dichotomization," *Applied Psychological Measurement*, 7 (Summer), 249-53.

Cox, Eli P. (1980), "The Optimal Number of Response Alternatives for a Scale: A Review," *Journal of Marketing Research*, 17 (November), 407-22

Devlin, Susan J. and H.K. Dong (1990), "Measuring the Quality of Service from the Customer's Perspective," *Proceedings of the National Communications Forum*. Chicago, IL: National Communications Forum.

Crask, Melvin R. and Richard J. Fox (1987), "An Exploration of the Interval Properties of Three Commonly Used Marketing Research Scales: A Magnitude Estimation Approach," *Journal of the Marketing Research Society*, 29 (3), 317-39.

Gerbing, David W. and James C. Anderson (1988), "An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment," *Journal of Marketing Research*, 25 (May), 186-92

Haley, Russell I. and Peter B. Case (1979), "Testing Thirteen Attitude Scales for Agreement and Brand Discrimination," *Journal of Marketing*, 42 (Fall), 20-32.

Hughes, John H. (1992), "Scale Tests for Support Services Measurements," Working Paper, Piscataway, NJ: Bellcore.

Jones, Lyle V. and Louis L. Thurstone (1955), "The Psychophysics of Semantics: An Experimental Investigation," *Journal of Applied Psychology*, 39 (1), 31-6.

Payne, Stanley L. (1951), *The Art of Asking Questions*. Princeton, NJ: Princeton University Press.

Singh, Jagdip (1990), "Identifying Consumer Dissatisfaction Response Styles: An Agenda for Future Research," *European Journal of Marketing*, 24 (6), 55-72.

Sudman, Seymour and Norman M. Bradburn (1982), *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossey-Bass.

TARP (1986), *Consumer Complaint Handling in America: An Update Study*. Washington, DC: White House Office of Consumer Affairs.

Van Heerden, J. and Joh Hoogstraten (1979), "Response Tendency in a Questionnaire Without Questions," *Applied Psychological Measurement*, 3 (Winter), 117-21.

Westbrook, Robert A. (1980), "A Rating Scale for Measuring Product/Service Satisfaction," *Journal of Marketing*, 44 (Fall), 68-72.

Susan J. Devlin is Director of Measurements Research at Bellcore, Piscataway, N.J.

H.K. Dong is Director of Learning Support at Bellcore, Piscataway, N.J.

Marbue Brown is a member of the Technical Staff, Measurements Research, Bellcore, Piscataway, N.J.