

Best Practices in the Assessment of Hotel-guest Attitudes

For guest surveys to provide useful information, those surveys must be properly designed.

BY MATTHEW SCHALL

It is gospel in the hospitality industry that guest satisfaction is essential to sustaining revenues and profits. Hampton Inn's web site, for instance, quotes Phil Cordell, the senior vice president of brand management, as saying "the quality and consistency of our guests' hotel experience... makes a tremendous difference in the success of our hotels." In their company's 1997 annual report, Wayne Huizenga, chairman of Extended Stay America, and CEO George D. Johnson, Jr., put the argument this way: "Superior satisfaction ratings and very high levels of intent to return and recommend to others (are) easily the best way to obtain new customers, and to keep the [current] ones returning again and again."

The effect of satisfaction on guests' intent to return is clear. One study, performed for Gaylord Opryland Nashville, found

a correlation of 0.76 between overall satisfaction and intent to return for 527 guests who responded to satisfaction surveys, indicating that satisfied guests are more likely to return than those who were somehow disappointed. Gaylord Opryland Nashville is far from the typical hotel property, given its size and facilities, and its managers actively track and respond to guest perceptions and staff attitudes.¹

¹ Because of the nature of Gaylord Opryland Nashville, these results might be different for other hotels. Although the property has been expanded several times since the following articles were published (and Gaylord now operates additional destination properties similar to Nashville's Opryland), the nature of the operation is explained in: Michael R. Evans and Robert D. Reid, "Opryland Hotel: Managing Nashville's Complete Destination," *Cornell Hotel and Restaurant Administration Quarterly*, Vol. 25, No. 4 (February 1985), pp. 44-55; and Marc Clark, "Training for Tradition at Opryland Hotel," *Cornell Hotel and Restaurant Administration Quarterly*, Vol. 31, No. 4 (February 1991), pp. 46-51.

© 2003, CORNELL UNIVERSITY

Improving customer satisfaction is a critical component of the hospitality industry's value proposition to guests.² Along that line, the hospitality industry spends millions of dollars per year to better assess guest satisfaction and understand the elements of loyalty. With an understanding of what causes guests to stay and to return, hoteliers can act to increase loyalty and improve image. Without this understanding, on the other hand, it is difficult to design programs that capture or retain guests and increase profits.

Inappropriate survey timing, question order, and—perhaps most important—sample size each can interfere with a survey's validity.

To identify the operational and marketing issues that influence satisfaction and loyalty, most hotels survey their guests in some way. In this article I explain how my associates and I design, administer, and interpret such questionnaires. Too often, hoteliers use questions, scales, and methods that they have adopted (or purchased) without careful consideration and thoughtful evaluation. Such instruments may have been designed for a different purpose, be unclear, use inappropriate scales, fail to make a valid measurement of key attitudes (such as satisfaction and loyalty), or provide limited knowledge about guest attitudes. Furthermore, just as poor questions generate misleading data, so do poor survey methods. Inappropriate survey timing, question order, and—perhaps most important—sample size each can interfere with a survey's validity.

The collection of valid and timely data is the key to gaining the information that hoteliers use to guide profitable decisions. In this context, evaluation of image at the brand level and performance at the property level is governed by the same rules. This article describes best practices for measuring guest satisfaction and loyalty. I discuss the best practices that my firm uses in

² See, for instance: Judy A. Siguaw and Cathy A. Enz, "Best Practices in Hotel Operations," *Cornell Hotel and Restaurant Administration Quarterly*, Vol. 40, No. 6 (December 1999), pp. 42–53.

survey design, including intent, clarity, scaling, and validity, as well as the methodological issues of timing, question order, and sample size.

Source note. The research studies and results presented here are the result of UniFocus LLP internal analyses of data from guest-satisfaction and -loyalty surveys at client hotels.³ Clients include single- to multi-hundred-property management companies, chains, hotel groups, and individual hotels in all segments of the hotel industry. All hotels on which analyses are performed are located in North America, South America, or Europe. I make assertions in this article that reflect the client company's practical experience.

Methodological Review

After a review of the issues involved in accurately measuring guest opinion, this paper examines methodological issues and then discusses examples and consequences of poorly performed research.

Validity. Validity refers to whether a question or survey measures the desired topic. Surprisingly, many of the questions used in hospitality-industry surveys are invalid. Poor measurement is a waste of the money spent on evaluation, and even worse, can lead to incorrect conclusions and costly mistakes. There are many types and measures of validity.⁴ This paper addresses aspects of validity that are critical to the hospitality industry.

Intent of the question. Intent in the context of validity refers to whether the question actually measures what it is intended to measure. This concept is sometimes called face validity, which can be expressed as follows: does the question read as though it measures what it is supposed to? An example of an item with an unclear intent is "Rate the condition of your room's carpet." Is the purpose of this question to measure (a) service (that is, housekeeping), (b) quality and standards assurance, or (c) assessment of the property's FF&E choices? If the question is about the cleanliness of the room's carpet, then it should be worded thus: "Was your room's carpet clean?"

³ Most assertions in this article that are not the author's are taken from the Sage Press University Paper Series. These volumes are useful references for non-academics and those academics who are not quantitatively oriented.

⁴ Edward G. Carmines, and Richard A. Zeller, *Reliability and Validity Assessment* (Newbury Park, CA: Sage Publications, V17, 1979), pp. 9–13.

If the goal involves quality assurance, the question is misplaced because it is not the guest's responsibility to judge whether the hotel is meeting its standards. Similarly, if the goal is to assess the quality of or need for FF&E expenditures, guests should not be involved. A walk through the rooms will accomplish the same goal without focusing guests' attention on something that may need improvement. Thus, this item is an example of a question that can lead one to the wrong conclusions. If guests answer with high scores because they are evaluating "carpet cleanliness," but the hotel's intent was to assess the room's décor or appearance, the property managers could reach conclusions about appearance based on housekeeping performance.

Operations versus attitudes. My firm's experience suggests that there are two areas where guests' opinions are critical to a property's success. One is operational (that is, making an evaluation of the hotel's functioning) and the other is attitudinal (that is, what the guest thinks of the stay and the property). This distinction is central to the application of survey results. Guests' assessment of operational matters, like whether the room was clean or the food of good quality, guide specific tactical actions by the management team. Guests' attitudes about those same factors, (e.g., cleanliness, quality, and other operational issues) contribute to overall feelings of satisfaction, intent to return, and value.

The importance of separating guests' operational assessments from attitudes is considerable, because the connection between operational characteristics and guest attitudes is not perfect. What is clean to one guest may be marginal or unacceptable to another. A guest who is loyal because of a rewards program may be willing to partially discount the importance of some operational issues and continue to stay at the hotel. Since the connection between operational considerations and the attitudes that drive guests' intent to return are not one-to-one, it is necessary to measure both. To avoid confusing the guest about a question's intent, one should separate a survey's operational and attitudinal questions.

Clarity of the question. One of the biggest problems with question clarity is the use of compound or "double barreled" questions that ask

more than one thing at a time.⁵ When one asks about two different matters in a single survey question, the response cannot be interpreted accurately. Consider the following example, which often appears on a hotel's guest-survey card: "Front-desk staff was friendly and efficient." This asks two different questions—one about friendliness, and the other about efficiency. It is quite possible that a front-desk employee can be friendly, but inefficient at a specific task, say, changing a room. Another may change the guest's room with great efficiency, but be so crisp about it as to seem gruff. No matter how the guest responds to the question, there is no way to tell what the result means. The logical fix for double-barreled items is to ask two separate questions—for instance, one question that asks whether the front-desk staff was friendly and another that asks whether those employees were efficient.

Even when the question focuses on a single topic, providing a clear context or instructions may enhance the clarity of the question. Consider the question, "How do you rate the quality of the food served to you?" Quality is an ambiguous term, and, indeed, the guest can use any criterion for her answer, including that she just didn't care for a particular dish. Clarity may be enhanced, however, when the guest is given instructions on answering the question. In this example, the instructions might include just a set of parenthesis that define quality for the respondent, or actually include a statement like "when answering the quality-of-food question, please think of the taste, temperature, and appearance."⁶

Unidimensionality. Unidimensionality is a statistical term that is conceptually similar to question clarity, in that unidimensional survey questions ask about just one topic at a time. When the responses to the questions on one topic are added together to form a single score, all the questions must assess the same thing, or the score's meaning may be unclear.⁷ This does not

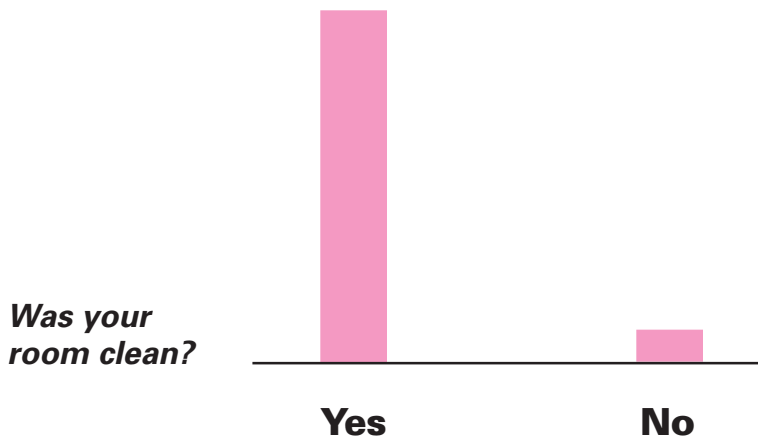
⁵ Jean M. Converse and Stanley Presser, *Survey Questions* (Newbury Park, CA: Sage Publications, V63, 1986), p. 13.

⁶ For guidance and resources for writing question instructions, see: *Ibid.*, pp. 18–31.

⁷ Edward G. Carmines and Richard A. Zeller, *Reliability and Validity Assessment* (Newbury Park, CA: Sage Publications, V17, 1979), pp. 59–70.

EXHIBIT 1

Response distribution to a two-point room-cleanliness question



mean that a survey cannot measure two or more different guest attitudes, as long as they are tallied separately. It is perfectly reasonable, for example, to evaluate both satisfaction with operational issues and feelings of loyalty in the same survey, but the resulting scores need to be discussed separately for clarity. Sometimes people raise the widely used Scholastic Aptitude Test as an example of an instrument where different scores, in this case verbal and quantitative, are added together, in apparent contradiction of the advice that scores be considered separately.⁸ Both the verbal and quantitative factors are themselves measures of a second-order or super factor commonly referred to as scholastic aptitude. In this case, both the factors measure the same thing (albeit with different types of questions), and adding the verbal and quantitative score together to form an overall scholastic ability score is acceptable.

A hospitality example associated with the corresponding dimensionality analysis is presented later in Exhibit 8. In this example, a fragment of an actual factor analysis shows measures of front desk and restaurant or culinary performance that define separate factors.

⁸ Commonly known as the SAT, published by the Educational Testing Service, Princeton, New Jersey.

Any items that are averaged or added together must correlate with each other. The typical approach to assessing unidimensionality involves the use of a statistical tool like factor analysis.⁹ Properly performed, this analysis identifies all the questions that measure the same thing.¹⁰ Once this analysis is complete, the scores for questions that measure the same factor may be added or averaged together to give a single number that represents the topic. Unidimensionality is crucial in hospitality-industry surveys, because many companies award bonuses to managers based on scores that are averages of guests' responses to survey questions. If the questions added together do not measure one single underlying factor, the bonuses are based on an invalid score.

Comparability of scales. Bonuses can, however, be based on a variety of scales, as long as those scales are measured individually. If different scores are added together, one not only needs to assess dimensionality, but also the comparability of the numbers being used. A GM, for example, may receive a bonus based on financial performance and guest-satisfaction scores. Usually, there is a weighting associated with such a bonus. For example, 80 percent of the bonus may be based on financial results and 20 percent on guest satisfaction. In this case, the bonus money is split into two separate funds, each of which is awarded according to the appropriate score. While that approach is perfectly valid, it is inappropriate to calculate a bonus based on the sum of financial-performance and guest-satisfaction measures. The reasons for this are that revenue-performance and guest-satisfaction scores are most likely unrelated. Further, financial performance is measured in numbers ranging from thousands to millions of dollars, while guest satisfaction is usually presented on a scale of 0 to 100. If these two numbers were simply added together, the financial score would overwhelm the guest-satisfaction scores. More formally, the weighted effect of the revenue score is so much

⁹ For a detailed discussion of factor analysis, see: Robert C. Lewis, "Isolating Differences in Hotel Attributes," *Cornell Hotel and Restaurant Administration Quarterly*, Vol. 25, No. 3 (November 1984), pp. 64–77.

¹⁰ Paul Spector, *Summated Rating Scale Construction* (Newbury Park, CA: Sage Publications, V82, 1992), pp. 13–17.

greater than that of guest satisfaction that guest satisfaction will not contribute enough points to a summed score to have an appreciable effect on the bonus, unless the scores are treated separately.

Applying factor analysis. My firm has identified the key guest-attitude factors based on results from over 300 factor analyses on answers to thousands of survey questions administered at hundreds of properties. The goal was to uncover the crucial factors that underlie guest attitudes and to identify a small number of questions that would best measure those factors. These analyses were undertaken because asking unnecessary questions raises survey costs, increases the length of the survey (which decreases response rates), and distracts the user of the data from focusing on what is essential. The research indicates that guest reports of satisfaction or dissatisfaction, positive or negative experiences, and loyalty (as gauged by intent to return) can be assessed by testing three main factors, namely, room, food, and staff.

Room. Key elements for the room are cleanliness, functioning amenities, and comfort. For a four-star airport hotel that caters to business travelers, for instance, the guest might focus on the comfort of the room's work space. For an economy property, comfort might mean a quiet, fresh-smelling room.

Food. Food experience has three core components—namely, the overall perception of quality, including taste, appearance, and temperature; the promptness of service; and the accuracy of order fulfillment.

Staff. Guests evaluate hotel employees primarily on friendliness, helpfulness or accuracy, and promptness of service. Once again, one must avoid asking compound (i.e., “double barreled”) questions on these items by asking separate questions on each point.

I include this summary of the results of hundreds of in-house studies to help readers evaluate whether they are asking their guests the right questions. Other key questions or patterns of questions may work just as well, but this particular set of questions has performed well at the many properties where it has been used. Statistical tools help with question selection. The questions that best define a factor are the top choices

for inclusion on the questionnaire.¹¹ This research has been used to produce instruments that use just a few key questions to measure each factor, resulting in short surveys that ask only what is necessary. With good questions to measure each motivating factor, it becomes possible to predict the likelihood that a guest will return and identify crucial areas for management intervention. For example, if a particular department's employees have poor friendliness scores, the property's managers can take effective action to increase guest loyalty by providing additional training for that department.

Scaling

In this context, the term “scale” has two meanings. First, it is the ruler used to measure a response, as when a question uses a seven-point Likert-type scale that might range from “very little agreement” to “very much agreement.” This ruler is formally called a response scale. Second, a scale also refers to the questions used to measure something specific, as in a ten-question scale that measures extroversion.

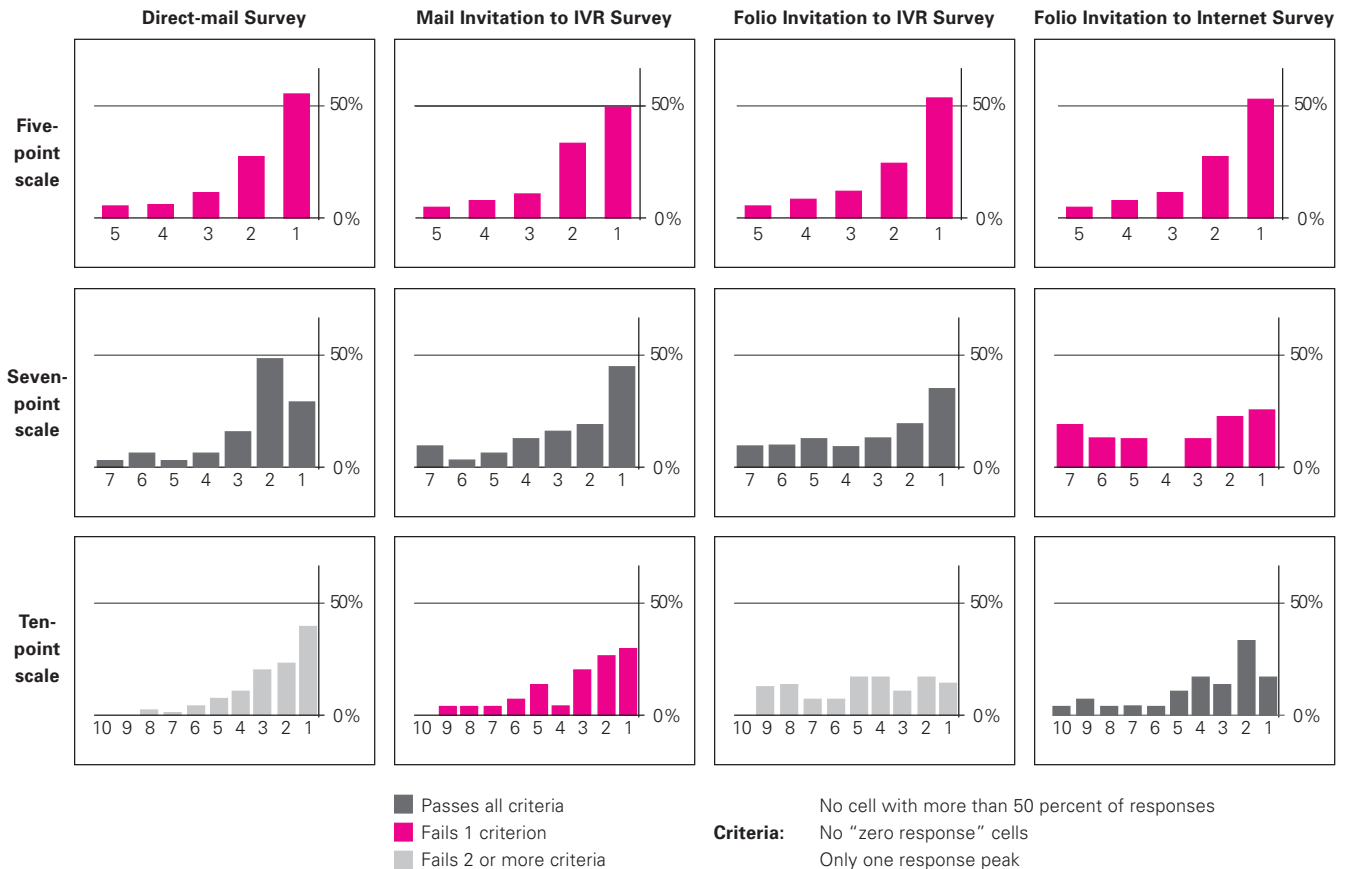
Inappropriate scaling or measurement of an item invalidates it. When my associates and I reviewed nearly two dozen surveys for this article, we found only three that included well-designed scales, while 17 contained double-barreled questions, asked imprecise or unclear questions such as “how do you rate...the bathroom,” or used scales that were biased toward positive responses.

There are three key characteristics to consider when scaling a question. First is the number of points or response options. For example, a scale with very good, good, fair, and poor has four points. Second is the wording of the scale's point values. Two issues are critical: the words used to describe the intervals or distances between the scale points and the language used on the end points, which are called scale anchors because they set the extreme values or anchors of the scale. Third is the presence of a midpoint or neutral point on the scale.

¹¹Jae-On Kim and Charles W. Mueller, *Introduction to Factor Analysis* (Newbury Park, CA: Sage Publications, Vol. 13, 1978), p. 24.

EXHIBIT 2

Response distributions by scale size (using the value for price paid as an example)



Scale size. In the hospitality industry, our studies indicate that most guests give positive responses to survey questions. This causes problems for managers who truly would like to improve on their guest services. Suppose we have the following item, with a two-point response:

My room was clean: __ Yes; __ No

For an upscale property, over 90 percent of all guests will answer yes—even those who are indifferent but did not find the room dirty. The response distribution of such a question is a bar chart like that shown in Exhibit 1 (see page 54).

This result conveys little useful information other than that a small percentage of the guests felt strongly enough to answer no. It does not tell how many guests were indifferent about their

room’s cleanliness, or thought that the room was clean enough, but could have been cleaner. Because of the two-point scale choice (yes or no), the question delivers poor-quality information.

One way to assess the best scale size for hospitality companies involves looking at the distributions of the responses to a question. Scale sizes can be evaluated by expressing the scores as a bar chart. The patterns for scales that are too large, too small, and about right, following this methodology, are:

- With too large a scale, that is, one with too many response options, people have a difficult time deciding which option to mark. This sometimes shows up in the data as responses that fall into groups and gaps, leav-

ing “holes” in the data. For example, suppose a survey asked for ratings on a 1-to-20 scale. Chances are that people would simplify such an overwhelming scale by responding in a pattern that effectively reduces a 20-point scale to one that has only three or four points.

- With too short a scale, on the other hand, many people tend to elect a single response, resulting in the condensing of responses to the point where information is lost, as in the case of Exhibit 1’s yes–no bar chart. One way to identify when a scale is too short is to look for a majority of the scores showing up on one point.
- An appropriate response scale usually has a distribution of scores that fills in all the responses, but does not over-emphasize one response. These right-size scales tend to have one peak in their distribution, that is, they are unimodal.

Through exhaustive study, my firm has determined that a seven-point scale is the optimum size for hospitality-industry questionnaires. To evaluate this matter, we undertook several studies to identify the scale that works best for the way guests answer questions about hotels. Studies were performed on mid- and large-size midprice and upscale hotels (defined using the Smith Travel Accommodations Reports criteria).

For this purpose, we compared five-, seven-, and ten-point scales. All surveys were administered with the same questions at the same properties at the same time. The six issues we examined were: room cleanliness, whether the respondent would recommend this hotel to others, overall satisfaction with stay, satisfaction with service, overall hotel condition, and value for the price paid. Four different invitation and survey-delivery methods were used—those being direct mail, a mailed invitation to an interactive voice response (IVR) survey, a folio invitation to an IVR, and a folio invitation to an internet survey.

Response distributions were compared for each scale size on all six questions, looking for the scale size that produced the “best” response distributions according to three criteria. Those criteria were a unimodal distribution and scores sufficiently spread out to convey information about people who have both moderate and ex-

EXHIBIT 3

Sample scale anchors

Very...	Very...
...Satisfied	...Dissatisfied
...Much agree	...Much disagree
...Positive	...Negative
...Valuable	...Costly
...Enjoyable	...Unpleasant
...Friendly	...Unfriendly
...Helpful	...Unhelpful

trema responses, but also no gaps or holes in the responses. To control for too few responses, we evaluated whether any single category drew more than half of the responses. At the other end, we judged whether the question had too large a response scale by whether some scale items drew no responses.

About right. Exhibit 2 shows a graph illustrating the study results using the value for price paid as an example. The distributions that met all three criteria are dark gray, those that meet only two out of three criteria are in color, and those that meet only one or none of the criteria are in light gray. An examination of Exhibit 2 shows that:

- In all of the five-point scales, one category drew over 50 percent of responses.
- Three of four ten-point scales have no responses at “10,” creating a hole in the data.
- The ten-point folio invitation to an IVR distribution resulted in three peaks, indicating that people were really using the scale to answer high, medium, and low (that is, the scale was too spread out for the question).
- The ten-point direct-mail distribution has several zero-response categories.
- Only one of the seven-point scales had a category with no responses.
- Three seven-point scales met all three criteria.
- These findings were consistent for all six questions, with the seven-point scale performing best in all cases.

As a consequence, based on those results and other studies, we recommend the use of a seven-point scale for hospitality surveys.

EXHIBIT 4

Scale anchors compared

Scale A							
Very Good						Very Poor	
7	6	5	4	3	2	1	N/A
Scale B							
Excellent						Very Poor	
7	6	5	4	3	2	1	N/A
Scale C							
Outstanding						Disappointing	
7	6	5	4	3	2	1	N/A

Many companies perform statistical analyses on scale-score results. Common analyses include correlations, *t*-tests, ANOVA, and regression. All of the common versions of these statistical tests require normal distributions (commonly known as bell curves), except the *t*-test, which uses symmetric distributions that look somewhat similar to normal distributions. The criteria of no holes in the response data, unimodal distributions, and no category with more than 50 percent of the responses correspond to characteristics of normal or *t*-distributions. While meeting those criteria does not guarantee that a statistical test is valid, if they are not met, many common statistical tests cannot be used.

Scale wording. The words used to describe the extreme points of a scale (i.e., the anchors) must also be carefully chosen. The key point regarding scale anchors is that they must be logical opposites. Moreover, they must make sense in consideration of the question’s context and intent. Typical oppositional anchors are shown in Exhibit 3.

To get a sense of how various anchors can be used, consider the seven-point scales shown in Exhibit 4. The only differences among those scales are the words anchoring the scales, that is, at the extremes. The anchors in Scale A are polar opposites, but there are times when a hospitality-industry survey would be more effective with an asymmetric pair of anchors—albeit still polar opposites, as in the case of Scale B. An example is the evaluation of hotel GMs for bonuses. “Very good” is often not considered rigorous enough

to meet the expectations of the brand image, and excellent is used instead. That is, the bar for evaluating GM performance is moved up from “very good” to “excellent” in keeping with the hotel’s standards of luxury and service.

Taking that point a step further, to evaluate the emotional components of a stay or a meal at a truly luxurious property, conventional words like “good” or even “excellent,” can seem trite, and they may not capture the distinctive nature of the experience. In this case, scales with anchors like those found in C may be used. Using unusual scale anchors like “outstanding” and “disappointing” signals the respondent that something outside normal experience is being evaluated. More specifically, such anchors acknowledge the guest’s high expectations for the experience. At the luxury level, even something that is very good may not be good enough; the experience must be extraordinary to make an impression.

Equal intervals. Consider a scale with the response points or scale of very good, good, fair, and poor, with the corresponding numbers of 4, 3, 2 and 1. Is the perceived or cognitive distance between “very good” and “good” the same as the distance between “fair” and “poor”? I argue that it is not. Fair is neutral or positive (as in fair weather) while poor is clearly negative, which creates a much larger gap between marking a 2 (fair) or a 1 (poor) than the cognitive distance between very good and good, which are both positive responses. If the respective numeric values associated with those words are 4, 3, 2, and 1, any mathematical and most statistical analyses of the data are invalid.¹² Examination of means and statistical tests such as factor analysis require that the responses be measured on an equal-interval scale. That is, the perceptual distances between the points must be of the same size.

Using “not applicable,” or N/A. The not-applicable response serves several critical functions. First, it gives guests an easy way to answer a question that does not apply to them. An example would be the question “Promptness of laundry delivery” when the guest did not use the laundry service. Making it easy for guests to re-

¹² Milton Lodge, *Magnitude Scaling* (Newbury Park, CA: Sage Publications, V25, 1981), pp. 8–25, 50–72.

spond to questions that do not apply to them decreases respondent frustration, because guests do not have to decide whether they should mark a positive, negative, or neutral response. Second, when a survey allows N/A responses and the guest leaves the response blank, one can assume that the respondent has no strong feelings one way or the other, and that a neutral response can be substituted for the blank response.

Using a neutral point. Another feature of the four-point scale just discussed is that it has no neutral or midpoint. This difference is highlighted in the three scales shown in Exhibit 5.

The four-point scale A has no neutral point. A scale of this kind diminishes response rates. As noted by my firm, guests typically fail to respond to 10 to 40 percent of the questions on a generic hospitality-industry questionnaire and that the rate of skipped questions increases when a neutral response is not available. In addition, a neutral response provides a way to treat missing data. When a guest skips a question, one can interpolate a neutral response for the missing values as long as “not applicable” is included as an option. Rather than lose the data for a question entirely, substituting a neutral for a blank greatly improves the ability to extract all available information from the survey data.¹³

Non-response is a serious problem, because it causes a loss of potentially valuable data. Several steps can reduce the number of non-responses to survey items. First, questions and their context must be made clear, thereby decreasing respondent frustration and reducing the time needed to complete the survey. Second, one can perform a skip analysis and then rank order questions according to the likelihood of their being skipped (with most-skipped questions last). A skip-response study, as performed at my firm, is an examination of the proportion of skipped responses for each item, followed by an evaluation of the proportion of skipped items once respondents have stopped answering. In one study, my associates and I found that once guests skipped two questions, there was better than a 50-percent chance that they would skip all remain-

EXHIBIT 5

Sample scales with and without a midpoint

Scale A	Excellent	Good	Fair	Poor	N/A					
Scale B	Excellent	5	4	3	2	1	Poor	N/A		
Scale C	Excellent	7	6	5	4	3	2	1	Very Poor	N/A

ing questions. Third, a neutral response—and usually a not-applicable (N/A) response, as well—are included to avoid forcing a positive or negative response that would encourage the guests to skip questions when they do not apply to the guest’s experience. If a noncommittal response is not available, respondents tend to skip the question. Even if the respondents want to provide an answer, they have to spend time deciding whether they want to give a positive or negative response to a question about a hotel service that they did not use. The lack of a not-applicable response decreases the guest’s ability to focus on the questions that do measure satisfaction and loyalty, and the resulting data are suspect.

Four-point scale. Scale A in Exhibit 5 is an attempt to avoid the problem of extreme scores that can afflict five-point scales (as mentioned above). In an attempt to reduce the number of extreme positive scores some survey authors use the four-point scale with two positive, one neutral or vaguely positive, and only one negative response. However, this scale does not, in fact, correct the bias toward positive responses as well as Scale B does with its five points (Exhibit 5). Instead, by using only one negative point, Scale A’s responses are biased toward the positive. Such a bias means that the users of this scale (e.g., hotel managers) are not collecting the negative feedback necessary to fix problems; instead, they are likely to get artificially favorable scores.

Five-point scale. Scale B has the advantage of being balanced—that is, it has two positive, two negative, and one neutral response. A scale balanced in this way produces more-accurate responses than the positively biased Scale A, as discussed with regard to equal intervals and positive bias. Since my firm’s research indicates that five-point scales still have a tendency toward extreme outcomes, however, Scale C remains the preferred choice.

¹³ John P. McIver and Edward G. Carmines, *Unidimensional Scaling* (Newbury Park, CA: Sage Publications, 1981), pp. 64–65.

The Effect of Time Lags on Guest Response

My associates and I administered a survey using two different invitation methods, namely, handing a survey card to the guest at checkout or sending a mailed invitation to take the same survey using an interactive-voice-response (IVR) system. The survey cards are a way to ask guests to respond to the survey immediately, while they are still physically in the hotel. The mail invitation incorporated a two-week delay for the invitation, plus any additional delay before respondents called the IVR.

In the matrix of responses, each cell (immediate card versus delayed mail invite to IVR) related to 18 to 36 properties, about 25 percent of which were budget or economy, 25 percent were upscale, and 50 percent were midscale. Response rates averaged about 10 percent with both survey methods.

Two *t*-tests were performed to evaluate the differences in responses for immediate comment card versus the delayed IVR. The *t*-tests were used to evaluate whether survey timing and administration method created differences in responses to two critical questions—to wit, the respondent's overall satisfaction and whether the respondent would recommend the hotel.

With regard to recommending the hotel, the result was $t(338) = 3.88$, $p < .05$, indicating a significant effect of delay (see next section for qualifications of this and the next statistic). For overall satisfaction, the result was $t(368) = 2.28$, $p < .05$, again showing a significant effect due to delay. As shown in the exhibit, guests gave significantly more positive scores on the delayed survey.

There is one flaw with this study, however. The response rates and sample sizes for the mail invitation to IVR were low, about 0.2 percent. The results from the statistical tests, while significant, may not be valid in that the population of guests may not be adequately represented in the IVR cells. The results are presented here for discussion and thought, but strong conclusions are not warranted. The tiny sample size may account for the fact that while the "recommend" scores are higher than the satisfaction scores, where the satisfaction scores are below neutral, the "recommend" scores are above neutral. There were 107 mail IVR responses, and 2,149 responses to the card surveys handed out at the front desk on checkout.

Note that some would argue that there is a positive bias introduced by the demand characteristics of handing a guest a survey. The argument would be that answering negatively is more difficult when facing a hotel employee. If such a demand characteristic was present in this study, then the effect of delay on responses was an even more powerful bias, since delay made responses more positive, unless the small sample size in the IVR cell has caused invalid results (which is quite possible).

Regardless, the message to the hospitality industry is that a time delay may make a significant difference to the accuracy of measurement for some critical guest attitudes. Those who delay measurement too long may be getting overly optimistic ratings of their properties.—M.S.

Effects of delay on survey responses:

<u>Invitation method</u>	<u>Comment card (at hotel)</u>	<u>Mail and IVR (delayed)</u>
Recommend hotel (7-point scale)	4.45 (n = 195)	5.03 (n = 385)
Overall satisfaction (5-point scale)	1.96 (n = 198)	2.38 (n = 388)

Notes: Willingness to recommend the hotel was judged on a seven-point scale (where seven was high), while overall satisfaction was rated on a five-point scale (where five was high).

Seven-point scale. As explained above, a seven-point scale like C should collect the best possible data. That scale is balanced with equal numbers of positive and negative responses, and it has a neutral point. Scale C also has enough spread to differentiate extremely positive or negative responses from those that are somewhat positive or negative. Additionally, it offers a "not-applicable" response, which is off the scale, as discussed earlier.

Survey Size

Determining the appropriate number of questions to include on a survey requires a delicate balance between keeping the survey short enough so that guests answer all the questions and making it long enough to gather the necessary information. Two well-researched and competing influences affect a respondent's decision to complete a survey. The first is the perceived amount of effort involved in filling out the survey.¹⁴ The longer the survey, the more time and effort is required of participants to complete the survey. That time and effort are seen as costs that may not be offset by the perceived value of providing responses. Counteracting the questionable utility of spending time on a survey is the salience of the survey to the respondent.¹⁵

¹⁴ Kim Sheehan, "E-mail Survey Response Rates: A Review," *Journal of Computer Mediated Communication*, Vol. 6, No. 2 (2001), p. 4 (as viewed at www.ascusc.org/jcmc/vol6/Issue2/sheehan.html).

¹⁵ Laura Branden, R. Mark Gritz, and Michael R. Pergamit, *The Effect of Interview Length on Attrition in the National Longitudinal Survey of Youth* (Washington, DC: U.S. Bureau of Labor Statistics, Amrch 1995 (as viewed at: www.bls.gov/ore/abstract/nl/nl950030.htm)).

The more salient or important the topic, the greater the likelihood that respondents will complete a lengthy survey.

Many surveys use several questions to measure a single attitude. For example, satisfaction with the room is often measured with questions about room cleanliness, décor, temperature, or appearance; whether the bed was comfortable; and whether everything was working. While multiple measures aid in identifying how the guest feels about the room, they also make the survey longer. My firm’s approach is to run a pilot test of a survey at a hotel (that is, preadminister the survey) and then correlate the items’ responses to each other. Any items with high correlations, usually .9 or greater, are considered to measure the same thing to such an extent that the questions are redundant, and only one of the highly correlated questions remained on the final survey.

In general, it’s best to keep surveys short provided that the minimally necessary amount of data is collected. In addition, short surveys may cost less to develop and be easier to administer.

Timing the Survey

Most people in the hospitality industry have interacted with frequent business travelers who ask what time zone, city, state, or hotel they are in. Given the importance of such repeat guests, a hotel needs to capture information from them immediately. One consultant who travels frequently for IBM, for instance, told me that he could not remember the characteristics of the hotel he had stayed at three days before: “All the hotels have just blurred together,” he said.

If hotel experiences blur together for frequent travelers, it becomes critically important that managers measure guest attitudes during or immediately after their stay. Otherwise, responses to questions about one hotel can be confused with attitudes about another. To evaluate the effect of the length of time between the guest’s stay and responses to survey questions, my associates and I undertook the study described in the accompanying sidebar (on the previous page).

Question Order

The sequence in which questions are asked can influence the way guests respond. Consider the two versions of a short guest survey found in

EXHIBIT 6

Question-order demonstration

	Very much agree			Very much disagree	
	5	4	3	2	1
Version 1: From general to specific					
Overall, my stay was very satisfying.	0	0	0	0	0
Employees were responsive to me.	0	0	0	0	0
My room was very clean.	0	0	0	0	0
I enjoyed the food and beverages.	0	0	0	0	0
Version 2: From specific to general					
Employees were responsive to me.	0	0	0	0	0
My room was very clean.	0	0	0	0	0
I enjoyed the food and beverages.	0	0	0	0	0
Overall, my stay was very satisfying.	0	0	0	0	0

Exhibit 6. They are identical except for the order in which the questions are asked.

In version one of Exhibit 6, guests are asked about their overall feelings first, and specifics second. If their answer to the overall-satisfaction question is positive, then the answers to the other questions are more likely to be positive, in a phenomenon known as the halo effect. The context of the first, all-encompassing question casts a shadow (e.g., a halo) over the subsequent questions.¹⁶ The second set of questions in Exhibit 6 asks the specific questions first, and then concludes with the summative overall-satisfaction question. Looked at another way, asking the general question last produces better data, with three questions influencing the answer to one, instead of one question influencing three responses.

Sample Sizes

Sample size is a critical issue in the hospitality industry. The results from any survey based on an inaccurate or too-small sample can be misleading, resulting in poor decisions and costly mistakes. At the property level, sample size refers to the number of guests or employees being surveyed. At the brand level, sample size can refer to the number of properties, rooms, guests, or surveys. Picking an appropriate sample size is

¹⁶ Edward L. Thorndike, “Constant Error in Psychological Ratings,” *Journal of Applied Psychology*, Vol. 4 (1920), pp. 25–29.

EXHIBIT 7

Appropriate sample sizes for different error rates and confidence levels

	Population size				
	1,000	2,000	3,000	4,000	5,000
3% error rate	<i>Sample sizes</i>				
90% confidence	431	549	604	636	657
95% confidence	516	696	787	842	879
99% confidence	648	959	1142	1262	1347
5% error rate					
90% confidence	214	240	250	255	258
95% confidence	278	322	341	350	357
99% confidence	399	498	543	569	586

EXHIBIT 8

Sample two-factor solution

Question	Factor 1: Food	Factor 2: Front desk
Food quality	.62	.09
Prompt restaurant service	.52	.08
Friendly restaurant service	.50	.08
Prompt check-in	.06	.59
Prompt check-out	.00	.57
Front-desk employees are friendly	.01	.56

a challenging task usually performed by a statistician, psychometrician, or other survey expert. However, the concept is relatively easy to understand and apply to hospitality-industry surveys.

Property-level samples. There are several issues in setting the correct sample size, including the desired error rate and confidence level.

Error rate refers to the precision of measurement. When we are told that election-poll results are accurate to, say, plus or minus 3 percent, they are describing the error rate for the results. If a candidate is projected to get 52 percent of the vote, the survey results are really saying that we can be relatively certain that the real number is between 49 and 55 percent (52 plus or minus 3 percentage points). For hospitality-industry surveys, the error rate must be small enough to identify whether a hotel’s performance on satisfaction and loyalty measures is decreasing—and the extent of any decline—so that appropriate action can be taken.

Confidence level refers to overall confidence in the results. Using the example above, if we set a .95 confidence level, it means that scores are expected to be within the error rate of what is true for the population 95 percent of the time. That is, if my hotel has a score of 80, a “95/3” means that my score will be within 3 percent of the true score 95 percent of the time if I were to measure repeatedly.

Population refers to the people of interest, in this case, hotel guests. One gives questionnaires to a sample, that is, a small group, of the members of a population to get a result that allows one to draw conclusions about the entire population’s attitudes. A hotel’s population is estimated as a function of the number of rooms, the average length of stay, and the occupancy rate, all for a given length of time—typically, one month. As an example, consider a hotel with 300 rooms, an average stay length of two days, and average occupancy rate of 66.7 percent, with one guest per room. The population of possible respondents in a 30-night month is calculated as follows: (Number of rooms) × (30 nights) × (Occupancy rate) ÷ (Average length of stay) or, in this case, (300 × 30 × .667) ÷ 2 = 3,001. That is, an estimated 3,000 guests who will stay in this hotel constitute each month’s survey population. Unlike many surveys, all hotel guests (that is, the

entire population) are invited to respond to a survey.

Response rate is the proportion of the population that responds to the questionnaire. Typically response rates vary from 0.3 to 30 percent.

Sample refers to the people from the population who are measured (i.e., those guests whose surveys are usable). In the case of hotel surveys, the sample is made up of those people who respond to the survey. Sample size is used two ways. First is the expected sample size that will result from launching a survey. That figure depends on the population size and the expected response rate. For example, if the expected response rate is 10 percent, the population is 3,000, and a survey is delivered to every possible respondent, then one would expect to see about 300 responses. The second application of sample size is the number of responses required to achieve a particular confidence level at a particular error rate. Exhibit 7 gives appropriate sample sizes for particular population sizes.

Certain statistical assumptions govern tables like the one in Exhibit 7. The figures given in the cells are minimum numbers that are chosen on the assumption that responses have a normal distribution and that the survey involved is relatively short, say, with just 5 to 15 questions. If the responses are not normally distributed, or there are more than 15 or so questions, the researcher should increase sample sizes by at least 50 percent above what's given in Exhibit 7. These calculations also assume that there are no skipped answers. If it is likely that a number of questionnaires will not be completed, the number of surveys distributed has to be increased to compensate (meaning that hotels must distribute the surveys during additional days or day parts).

In general, the more responses collected, the higher the confidence in the results (with the assumption that the sample is representative). This makes sense: if the population is 1,000 hotel guests, and we get 1,000 responses to our attitudinal survey, barring any inaccuracies in response, we then know what our entire population is thinking (i.e., 100-percent confidence). Accepting a confidence level of .95 or .90, for instance, will require fewer respondents, but will result in a decrease in accuracy.

Brand Sample Sizes

The questions asked on brand-level surveys are different from those asked on property-level surveys. Management companies are concerned with property and executive performance and brand loyalty. Performance is usually evaluated against a minimum set of standards or compares each hotel to the best-performing hotels in the system. A general manager may be told, for instance, that the system standard is 80 percent for guest satisfaction, but that his or her personal score is only 72 percent, or it may be that this GM's property has the second-best F&B-department score of all the properties in the brand. Sample size aside, there is business value in comparing numbers, and no statistical test is necessary, as long as the score is accurate.

Methodology Failures

Hospitality operators typically face two major methodological problems in trying to conduct or analyze surveys. One is a violation of the principle of unidimensionality, and the other is the difficulty in achieving an appropriate sample size.

Unidimensionality, violated. It seems that managers want to create a single overall score from a survey, which is typically calculated by averaging all guest responses. At this point, the unidimensionality of the questions on the survey is critical. As indicated above, it is not appropriate to combine scores of unrelated questions. As an example of when not to average scores, consider the fragment of a factor analysis performed on a set of questions that measure two different factors, as found in Exhibit 8.

It is clear from an examination of Exhibit 8 that there are two factors, one for food and related issues, and one for guests' treatment at the front desk. These are in no way related and cannot be summed or averaged in any way. The result would not be valid. Instead, two different scores must be formed, one for food and one for front-desk performance.

Inadequate sample size. People's livelihood often rides on survey results, but if the sample size is too small the results can easily be skewed in one direction or another. Here's how that might occur. Say that you are the GM of a 100-room midprice property with an 80-percent occupancy rate and an average guest stay of two

EXHIBIT 9

Hypothetical satisfaction scores

Typical scores (N = 25):

Score	Number of scores
7	8
6	10
5	2
4	2
3	1
2	1
1	1

Average is 5.6, or 80.0%

"Last month's" scores (N = 25):

Score	Number of scores
7	7
6	9
5	2
4	2
3	2
2	1
1	2

Average is 5.24, or 74.9%

EXHIBIT 10

Hypothetical satisfaction scores with an appropriate sample size

Typical scores (N = 250):

Score	Number of scores
7	80
6	100
5	20
4	20
3	10
2	10
1	10

Average is 5.6, or 80.0%

Two changes (N = 250):

Score	Number of scores
7	79
6	99
5	20
4	20
3	11
2	10
1	11

Average is 5.56, or 79.5%

Actual change (N = 250):

Score	Number of scores
7	70
6	90
5	20
4	20
3	20
2	10
1	20

Average is 5.24, or 74.9%

nights, or a monthly population of 1,200. Your bonus, promotions, and perhaps your job depend on your performance against a customer-satisfaction standard of 80 percent. The corporate office has retained a survey firm that makes 75 telephone calls and achieves 25 completed guest-satisfaction surveys each month. The chain's executives use this information to evaluate your performance against the 80-percent achievement standard. Based on your hotel's size and occupancy rate, interpolating between population sizes of 1,000 and 2,000, at a 90-percent confidence level with 5-percent error rate, the minimum sample size from examination of Exhibit 7 would be 220 completed surveys.

In a typical month, you see the results shown in the first table in Exhibit 9. When divided by the maximum possible score of 7.0, the average score of 5.6 calculated in Exhibit 9 comes out to an even 80-percent satisfaction rating. Say that trouble arises, however, when last month you saw the statistics in the second table in Exhibit 9, which results in an average score of 5.24, or just 74.9-percent satisfaction. Only two scores changed from the typical table to "last month's" table, one tally from a 7 to a 3, and the other from a 6 to a 1. With this change, you are now failing to meet the 80-percent goal.

The real concern here is the inadequate sample size used by the survey firm. Revisiting the calculations with a reasonable sample of 250 reveals the critical importance of choosing an appropriate sample size. The tables in Exhibit 10 show the same exercise performed with a sample size of 250 under three conditions. The first table ("typical") shows the property's usual scores from Exhibit 9. In the second column ("two changes"), two people have changed their ratings, as occurred in the second table of Exhibit 9 (that is, one from 7 to 3 and the other from 6 to 1). The third table ("actual change"), shows that 18 guests would have to shift the ratings from top to bottom to achieve the 8-percent change that so affected the small-sample results when just two people changed in Exhibit 9.

The logic here is that a decrease in scores from just two people is easily made up for with a slight increase in the next month's scores, as long as the sample size is adequate. However, if the decrease is real and not just a fluctuation caused by two

unhappy customers, bringing the hotel's performance back to 80 percent is much more difficult. With a too small sample size, on the other hand, there is no way to tell whether a score decrease is a one-month fluke or a major concern.

The key here is that with too small a sample size, small fluctuations make such a large difference that they can lead executives to make inappropriate decisions. With an appropriate sample size, the movement of two people (as shown in the "two changes" table in Exhibit 10) from high to low scores has little effect on the satisfaction score. However, if the sample size is appropriate and the results really indicate that 8 percent of the respondents switched from highly satisfied to dissatisfied, the management has a severe problem—which is that guest perceptions have changed and the satisfaction rating has dropped by over 5 percentage points (or, more than 6 percent).

Regional-executive issues. The above example describes the need for a sufficient sample size from the property's perspective. The same logic applies to achieving a sufficient sample size from each property in a region. Management companies compile data from many properties to examine performance by property type, by type of guest staying at the hotel (e.g., business, leisure, transient), and by region.

Consider a property-management company with 100 hotels with an even geographical distribution over four regions—ten luxury properties, ten resorts, and eighty upscale properties. Assume a mean of 100 rooms per hotel. If an average of 25 surveys are collected from each hotel each month (as often occurs), then over a three-month period, the chain will collect surveys from 750 luxury guests, 750 resort guests, and 6,000 upscale guests.

In the luxury market, assume that on average 30 percent of guests stay for business reasons, and the remaining 70 percent are leisure customers. Each month, the management company examines a score formed by averaging four key driver questions, overall satisfaction, overall experience, value for price, and intent to return. Call this the "Excellence" score. This score is used to evaluate regional performance and that of its executives, including GMs. Assume that these questions all indicate the same factor, and that it

is appropriate to add the scores together. When the property-management company wants to examine the Excellence score for business travelers (30 percent of guests) in the north region (25 percent of guests) who stay at luxury properties (750 responses), they are looking at only $.3 \times .25 \times 750$ or 56 surveys.

Needless to say, that sample size is tiny. For the ten luxury hotels with 100 rooms each, at an 80-percent occupancy and with an average stay of four nights, over a 90-day period the population is 180,000, and from Exhibit 7 it is certain that more than 600 surveys are required to produce an accurate estimate of the Excellence score—yet only 56 are being collected. In sum, the score that regional executives are being evaluated on is not precise because it is based on too few surveys. An executive should not have confidence in the score if only 25 surveys are collected from each property each month.

Evaluating Survey Information

A key to profitability in the hospitality industry is a clear understanding of what leads to satisfied and loyal guests. It's dangerous, however, to take action on misleading information regarding guest satisfaction and loyalty. This article is intended to provide hospitality practitioners with the knowledge needed to evaluate the quality of the information being supplied by research services, whether those services are in-house or being supplied by outside vendors. The key issues to examine with regard to surveys is whether (1) those surveys were constructed with valid questions that measure what they were intended to measure, (2) the survey's scale is appropriate, and (3) the sample sizes are sufficient to make reasonable conclusions about guests' attitudes. ■



Matthew Schall, Ph.D., is vice president, research and development for Carrollton, Texas-based UNIFocus (mschall@unifocus.com).