Getting the Truth into Workplace Surveys

by Palmer Morrel-Samuels

Harvard Business Review

February 2002

Getting the Truth into Workplace Surveys

by Palmer Morrel-Samuels

Many companies use surveys to measure employee motivation, job performance, and the effectiveness of HR programs. But more often than not, these tools miss important information and can even create new problems.

THERE'S NO DOUBT that companies can benefit from workplace surveys and questionnaires. A GTE survey in the mid-1990s, for example, revealed that the performance of its different billing operations, as measured by the accuracy of bills sent out, was closely tied to the leadership style of the unit managers. Units whose managers exercised a relatively high degree of control made more mistakes than units with more autonomous workforces. By encouraging changes in leadership style through training sessions, discussion groups, and videos, GTE was able to improve overall billing accuracy by 22% in the year following the survey and another 24% the year after.

Unfortunately, not all assessments produce such useful information, and some of the failures are spectacular. In 1997, for instance, United Parcel Service was hit by a costly strike just ten months after receiving impressive marks on its regular annual survey on worker morale. Although the survey had found that overall employee satisfaction was very high, it had failed to uncover bitter complaints about the proliferation of part-time jobs within the company, a central issue during the strike. In other cases where failure occurs, questionnaires themselves can cause the company's problems. Dayton Hudson Corporation, one of the nation's largest retailers, reached an out-of-court settlement with a group of employees who had won an injunction against the company's use of a standardized personality test that employees had viewed as an invasion of privacy.

What makes the difference between a good workplace survey and a bad one? The difference, quite simply, is careful and informed design. And it's an unfortunate truth that too many managers and HR professionals have fallen behind advances in survey design. Although the last decade has brought dramatic changes in the field and seen a fivefold increase in the number of publications describing survey results in corporations, many managers still apply design principles formulated 40 or 50 years ago.

In this article, we'll explore some of the more glaring failures in design and

provide 16 guidelines to help companies improve their workplace surveys. These guidelines are based on peer-reviewed research from education and the behavioral sciences, general knowledge in the field of survey design, and our company's experience designing and revising assessments for large corporations. Managers can use these rules either as a primer for developing their own questionnaires or as a reference to assess the quality of work they commission. These recommendations are not intended to serve as absolute rules. But applied judiciously, they will increase response rates and popular support along with accuracy and usefulness. Two years ago, International Truck and Engine Corporation (hereafter called "International") revised its annual workplace survey using our guidelines and saw a leap in the response rate from 33% to 66% of the workforce. These guidelines – and the problems they address – fall into five areas: content, format, language, measurement, and administration.

Guidelines for **Content**

1. Ask questions about observable behavior rather than thoughts or motives. Many surveys, particularly those designed to assess performance or leadership skill, ask respondents to speculate about the character traits or ideas of other individuals. Our recent work with Duke Energy's Talent Management Group, for example, showed that the working notes for a leadership assessment asked respondents to rate the extent to which their project leader "understands the business and the marketplace." Another question asked respondents to rate the person's ability to "think globally."

While interest in the answers to those questions is understandable, the company is unlikely to obtain the answers by asking the questions directly. For a start, the results of such opinion-based questions are too easy to dispute. Leaders whose understanding of the marketplace was criticized could quite reasonably argue that they understood the company's customers and market better

than the respondents imagined. More important, though, the responses to such questions are often biased by associations about the person being evaluated. For example, a substantial body of research shows that people with symmetrical faces, babyish facial features, and large eyes are often perceived to be relatively honest. Indeed, inferences based on appearance are remarkably common, as the prevalence of stereotypes suggests.

The best way around these problems is to ask questions about specific, observable behavior and let respondents draw on their own, firsthand, experience. This minimizes the potential for distortion. Referring again to the Duke Energy assessment, we revised the question on understanding the marketplace so that it asked respondents to estimate how often the leader "resolves complaints from customers quickly and thoroughly." Although the change did not completely remove the subjectivity of the evaluation – raters and leaders might disagree about what constitutes quick and thorough resolution – at least responses could be tied to discrete events and behaviors that could be tabulated, analyzed, and discussed.

2. Include some items that can be independently verified. Clearly, if there is no relation between survey responses and verifiable facts, something is amiss. Conversely, verifiable responses allow you to reach conclusions about the survey's validity, which is particularly important if the survey measures something new or unusual. For example, we formulated a customized 360-degree assessment tool to evaluate leadership skill at the technology services company EDS. In order to be sure that the test results were valid, we asked (among other validity checks) if the leader "establishes loyal and enduring relationships" with colleagues and staff; we then compared these scores with objective measures, such as staff retention data, from the leader's unit. The high correlation of these measures, along with oth-

ers, allowed us to prove the assessment's validity when we reported the results and claimed that the survey actually measured what it was designed to measure. In other assessments, we frequently also ask respondents to rate the profitability of their units, which we can then compare with actual profits.

In another case, we designed an anonymous skill assessment for the training department of one of the nation's largest

One of the best-documented response biases is the tendency of respondents to agree with questions, a tendency that becomes more pronounced as work progresses through a survey.

vehicle manufacturers and found that 76% of the engineers believed their skills were above the company average. Only 50% of any group can be above the average, of course, so the survey showed how far employee perceptions about this aspect of their work were out of step with reality. The results were invaluable for promoting enrollment in the company's voluntary training program, because few people could argue with the conclusion that 26% of the respondents – nearly 8,000 engineers – had a mistakenly favorable view of their skills.

In addition to posing questions with verifiable answers, asking qualitative questions in a quantitative survey, although counterintuitive, can provide a way to validate the results. In an employee survey we analyzed for EDS in 2000, we engaged independent, objective readers to classify the topic and valence (positive, negative, or neutral)

Palmer Morrel-Samuels is a research psychologist with more than 20 years of academic and consulting experience. A former research scientist at the University of Michigan Business School, in Ann Arbor, he is currently the president of Employee Motivation and Performance Assessment (surveysforbusiness.com), also in Ann Arbor.

of all written comments – 45,000 of them. We then examined the correlation between these classifications and the quantitative data contained in the survey ratings from all 66,000 respondents. The tight correlation between ratings and comments in each section of the survey – high ratings accompanying positive comments – gave us strong evidence of the survey's validity.

3. Measure only behaviors that have a recognized link to your company's performance. This rule may seem obvious, but as many as three-quarters of the questions (such as "I know about my company's new office of internal affairs") in surveys we review have no clear link to any business outcome or to job performance. This shortcoming explains many of the more startling survey failures. Most often, the problem arises because questions have not been systematically chosen. To avoid this, we use a two-step process to select question topics. First, we interview informed stakeholders, asking them to describe the main problems and what they think their causes are. Then, we review published research to identify known pairings of problems and causes.

For instance, to build a survey for International, we interviewed nearly 100 managers, employees, union representatives, and executives in the workforce of 18,000. We asked each to specify what aspect of performance they thought most needed improvement and what they believed was its primary cause. Interviewees all agreed that the defect rate required improvement but were less certain in identifying behaviors possibly causing the problem. Research on quality, however, seemed to confirm the suspicion of some stakeholders that improving communication would lower the defect rate.

As a result, we included a number of questions about communication in the survey. One question asked respondents to indicate how often "In our department, we receive all the information we need to get our jobs done." The results confirmed that poor communication was indeed associated with the defect rate. The company then implemented

a pilot program at one of its larger manufacturing facilities to improve communication within and between departments. Following this intervention, communication scores at the pilot site rose 9.5% while defects fell 19%. Although any of a number of factors may have been behind the defect rate, it was incontestable that the more communication improved, the more the defect rate fell.

Guidelines for *Format*

4. Keep sections of the survey unlabeled and uninterrupted by page breaks. Boxes, topic labels, and other innocuous-looking details on surveys can skew responses subtly and even substantially. The reason is relatively straightforward: As extensive research shows, respondents tend to respond similarly to questions they think relate to each other. Several years ago, we were asked to revise an employee questionnaire for a large parcel-delivery service based in Europe. The survey contained approximately 120 questions divided into 25 sections, with each section having its own label ("benefits," "communication," and so on) and set off in its own box. When we looked at the results, we spotted some unlikely correlations between average scores for certain sections and corresponding performance measures. For example, teamwork seemed to be negatively correlated with on-time delivery.

A statistical test revealed the source of the problem. Questions in some sections spanned two pages and therefore appeared in two separate boxes. Consequently, respondents treated the material in each box as if it addressed a separate topic. We solved the problem by simply removing the boxes, labels, and page breaks that interrupted some sections. The changes in formatting encouraged respondents to consider each question on its own merits; although the changes were subtle, they had a profound impact on the survey results.

5. Design sections to contain a similar number of items, and questions a similar number of words. Research and our own experience show that the more ques-

tions you ask, the higher the resulting scores for the entire section tend to be. Similarly, respondents often give higher ratings to questions that contain more words and require more time for reflection. Maintaining fairly equal question and section lengths provides the highest probability that you'll obtain compatible survey responses across all questions.

A customer satisfaction questionnaire used by a large retailer in the Northwest illustrates those dangers. In evaluating the survey, we found that longer questions and longer sections evoked higher ratings, regardless of the product being evaluated. Together, response biases produced by these two question characteristics elevated scores on the survey's final question ("How likely is it that you will repurchase from us?") and lowered the overall accuracy of the survey's findings. The company could have avoided both of these problems by maintaining consistent question and section length.

The same response bias – wherein scores increase with question and section length – will also elevate scores in excessively long surveys. In addition, the average score for survey questions increases as a respondent works through a questionnaire: It is not unusual to see the average score on a 100-question survey climb by 5%. At the same time, research and our experience show that the range of responses (the standard deviation) usually becomes smaller.

6. Place questions about respondent demographics last in employee surveys but first in performance appraisals. An optional section on demographics is a staple of customer questionnaires, and its value is uncontestable. Questions about demographics also frequently appear in employee surveys since managers believe the generated information can produce useful general data about workforce trends. Of course, it is imperative to avoid demographic questions that can seem invasive or irrelevant.

Including demographic questions, however, can dramatically depress employee response rates, especially when respondents feel that their anonymity may be jeopardized. A survey carried

out in 1999 by one of the nation's largest appliance manufacturers began by asking respondents whether they belonged to a union. Most of the union employees stopped filling out their surveys at this point; they reportedly feared that the data would be used to make misleading comparisons with unrepresented workers and that those comparisons could weaken the union's position during future contract negotiations.

In employee surveys, it's generally best to put demographic questions at the end, make them optional, and minimize their number. Such placement avoids creating an initial negative reaction at the very moment when readers are deciding whether to participate. A 1990 study by M. T. Roberson and E. Sundstrom found that moving demographic questions to the end of an employee survey improves response rates by around 8%.

In contrast to employee surveys, performance appraisals and leadership evaluations should include demographic questions and identifying items at the beginning. Placing those items first highlights their importance and increases the likelihood that respondents will answer them fully.

Guidelines for *Language*

7. Avoid terms that have strong associations. This rule of language is one of the most frequently ignored. Metaphor plays a prominent role in descriptions of management, but it can also trigger associations that bias responses. A leadership evaluation conducted in the mid-1990s by one of the nation's largest manufacturers of photographic equipment asked respondents whether their team leader "takes bold strides" and "has a strong grasp" of complicated issues. While such phrases are commonly used to describe leadership qualities, they are counterproductive in surveys because they can trigger associations favoring males, whose stride length and grip strength, on average, exceed those of women. As a result, the leadership ratings of male leaders for this assessment were unfairly elevated. Here, simple revisions in wording solved the problem: "Has a strong grasp of complex problems" was changed to "Discusses complex problems with precision and clarity." Subsequently, we found—as published research leads us to expect – no significant difference between the average scores of male and female leaders. We have observed similar results when words that trigger ethnic and religious associations have been changed.

8. Change the wording in about one-third of questions so that the desired answer is negative. One of the best-documented response biases is the tendency of respondents to agree with questions, a tendency that becomes more pronounced as work progresses through a survey. The best way to overcome this bias is to periodically introduce questions that are phrased negatively. It's possible to transform almost any question or statement ("In my department, we do a good job of resolving conflicts") to its opposite ("In my department, we do a poor job of resolving con-

---

Guidelines for Superior Survey Design

Content

1. Ask questions about observable behavior rather than thoughts or motives.
2. Include some items that can be independently verified.
3. Measure only behaviors that have a recognized link to your company's performance.

Format

4. Keep sections of the survey unlabeled and uninterrupted by page breaks.
5. Design sections to contain a similar number of items, and questions a similar number of words.
6. Place questions about respondent demographics last in employee surveys but first in performance appraisals.

Language

7. Avoid terms that have strong associations.
8. Change the wording in about one-third of questions so that the desired answer is negative.
9. Avoid merging two disconnected topics into one question.

Measurement

10. Create a response scale with numbers at regularly spaced intervals and words only at each end.
11. If possible, use a response scale that asks respondents to estimate a frequency.
12. Use only one response scale that offers an odd number of options.
13. Avoid questions that require rankings.

Administration

14. Make workplace surveys individually anonymous and demonstrate that they remain so.
15. In large organizations, make the department the primary unit of analysis for company surveys.
16. Make sure that employees can complete the survey in about 20 minutes.

flicts") without creating tortuous wording, double negatives, or the like. This practice is quite common. When airline personnel ask passengers about their baggage, they usually ask one question so the desired answer is yes and another so the answer is no. For instance, "Did you pack your bags yourself?" might be followed by "Have your bags been out of your control since they were packed?"

It is also important to describe reverse wording in the instructions to the survey and to clearly signal its presence to respondents. Readers can easily miss minor word changes; a statement such as "My leader makes unfair hiring decisions" might be misread as "My leader makes fair hiring decisions." So the wording of the negative questions must be carefully considered. One good way to prepare readers for this possibility within the questionnaire is to introduce a simple reversed item early on, in the third or fourth question. This reminds respondents about the presence of these kinds of queries throughout the survey. In our experience, we've found a good rule of thumb is to change the wording in about one-third of the questions.

9. Avoid merging two disconnected topics into one question. Many survey questions combine two elements. When items are associated, it makes sense to minimize the length of the survey by combining them, but at other times, merging two elements can be problematic. For example, a leadership assessment at a telecommunications company in the late 1990s asked employees to rate their leader's skill at "hiring staff and setting compensation." Clearly, data from such a question would result in little insight about a leader's specific skill in each of the two related but distinct tasks. In determining whether to include two related elements in the same question, decide whether the behaviors associated with them will require the same intervention if they need to be fixed. It can be quite reasonable to ask employees whether they think a leader both "provides and responds to constructive feedback" because both processes (to various degrees) require insight, tact, candor, flexibility, and a

willingness to learn. But asking about hiring and compensation at the same time will probably elicit muddied responses of little specific usefulness.

Guidelines for *Measurement*

10. Create a response scale with numbers at regularly spaced intervals and words only at each end. Many surveys invite respondents to evaluate an item by selecting words that best fit their own reactions. For instance, a global computer company's annual performance appraisal asked managers to evaluate employees by ticking one of five boxes labeled "unacceptable" to "far exceeds expectations." (See the top of the exhibit "Numbers Are Better than Words.")

The results of this kind of evaluation, however, are notoriously unreliable because they are influenced by a variety of extraneous factors. The biggest problem is that each response option on the scale contains different words, and so it is difficult to place the responses on an evenly spaced mathematical continuum in order to conduct statistical tests. Although the labels may be in a plausible order, the distance between each pair of classifications on the continuum remains unknown. For many people, for instance, "unacceptable" and "does not meet expectations" may be closer to each other than "meets expectations" and "exceeds expectations" are to each other. In addition, the response scale uses words that overlap ("exceeds" and "far exceeds") and that may mean different things to different people over time. Therefore, it is difficult to compare ratings on these scales from different managers in different years or to compare ratings from different departments, geographic regions, and even seasons.

You can avoid these and other distortions created by word labels by using a scale with only two word labels, one at either end with a range of numbers in between. Questions answered with numerical scales may not appear to be very different from those with word

answers, but the responses to them are far more reliable and can be submitted to a much more informative statistical analysis.

11. If possible, use a response scale that asks respondents to estimate a frequency. Relying on a numerical scale is only part of the story. There can still be a great deal of subjectivity in the question or in the words at each end of the scale that you'll need to eliminate. For instance, an employee survey we reviewed in the late 1990s asked respondents how much they agreed with the question: "Are you dedicated to quality in all that you do?" People were asked to tick a box on a scale between "disagree

Single-scale surveys take less time to complete, provide more reliable data, and make quantitative comparisons between different items much easier than multiple-scale surveys.

strongly" and "agree strongly." But questions that invite respondents to measure extent of agreement often produce biased responses. The bias may be especially pronounced if, as in our example above, disagreement would be unflattering to the respondent. After all, who would say that they were *not* dedicated to quality? Naturally, responses to this survey question were clustered at the high end of the scale.

The best way around the problem, we've found, is to invite respondents to provide an estimate of frequency, with percentages or ratings between "never" and "always," as shown in the lower part of the exhibit "Numbers Are Better than Words." For example, in conducting a nationwide benchmark survey of employee motivation, we asked: "What percent of the teams in your company produce high-quality work?" In contrast to the agree-disagree question on quality mentioned above, we used a rating scale with numbers and obtained a normal curve of responses (see the results for both types of surveys in the exhibit

"Well-Designed Surveys Produce Normal Results"), indicating that the responses were unbiased. What's more, a large body of research confirms that respondents' frequency estimations are typically quite reliable and accurate, even if they'd never consciously kept track of the behaviors examined in the survey.

12. Use only one response scale that offers an odd number of options. Many surveys have a jumble of different response scales, jumping from one to another without warning. A survey currently being used by a large hotel chain asks respondents to rate the service's friendliness on a scale from "very unfriendly" to "very friendly," then the service's efficiency on a scale from "very inefficient" to "very efficient," and so on for dozens of questions about the hotel's service. One response scale, such as "never" to "always" with numbered ratings in between, allows for an easy comparison of responses and is simpler for

respondents. Single-scale surveys take less time to complete, provide more reliable data, and make quantitative comparisons between different items much easier than multiple-scale surveys.

We find that it's advisable to provide an odd number of response alternatives, so that respondents have the option of registering a neutral opinion. We also advocate including a "don't know" or "not applicable" answer (preferably made to look different from the other answer options, as illustrated in the exhibit). Without that option, respondents may feel compelled to provide answers that they know are worthless. Including this option enhances response rates and makes it less likely that respondents will leave blanks or abandon the survey in the middle.

Take care not to offer too many or too few response options. In its annual employee survey, one of the nation's largest oil companies asks employees about attitudes and offers them only

two response alternatives: "agree" or "disagree." Inevitably, managers complain that the results are simplistic and difficult to interpret. We have found that a graded response scale with seven or 11 alternatives (the latter for scales from 0% to 100% in increments of ten) furnishes sufficiently detailed results.

13. Avoid questions that require rankings. Many surveys require respondents to rank a number of items in order of preference. A survey we reviewed in 1997 asked people to "Rank in ascending order of severity the problems threatening productivity in your department: on-the-job injuries, absenteeism, attrition, out-of-specification materials from vendors, lack of tools." Research shows, however, that responses to such questions are biased by a host of factors – most prominently the number, order, and selection of items. Respondents will best remember a list's first and last items and will tend to assign them the top and bottom ranks. Moreover, other research

Numbers Are Better than Words

Poorly designed surveys may contain questions that ask respondents to choose a word or phrase that fits their attitude. But the response options will mean different things to different people, and it will be difficult to apply informative statistical tests to the resulting data.

Please rate this employee's effectiveness in maintaining quality standards.				
◯	◯	◯	◯	◯
unacceptable	does not meet expectations	meets expectations	exceeds expectations	far exceeds expectations

Well-designed surveys ask respondents to select a numerical answer on a continuum between two well-understood word choices, preferably a continuum that requires estimating a frequency. The inclusion of a "don't know" or "not applicable" option enhances response rates because people know they can provide a truthful answer rather than just fill in a meaningless one.

During the last year, my department has produced high-quality work.							
◯	◯	◯	◯	◯	◯	◯	▭
1	2	3	4	5	6	7	
never						always	not applicable

shows that a ranking question can disrupt ratings on subsequent questions, presumably because respondents become sensitized to the topic of the ranking question.

*Guidelines for **Administration*** ⌐

14. Make workplace surveys individually anonymous and demonstrate that they remain so. As we have already pointed out, respondents are much more likely to participate in surveys if they are confident that personal anonymity is guaranteed. In our employee survey for International, we told employees that the anonymous surveys contained no hidden marks and that we would never be able to connect any individual survey to a specific employee. We backed up this claim by having boxes of spare surveys (under minimal supervision, to discourage people from submitting more than one questionnaire) at every facility. Access to all those loose surveys went a long way toward reassuring people about our commitment to anonymity.

The desire of respondents for anonymity explains why many companies prefer using paper-based surveys, even when all employees have access to a computer network. Most workers are savvy enough to know that each computer has a unique fingerprint and that passwords can be easily decrypted or overridden. A 2001 pilot test of a leadership assessment at Duke Energy illustrates the problems of administering surveys electronically. Duke ran, in parallel, an electronic and a paper-based version of its 360-degree leadership assessment so that the company could complete a cost-benefit analysis of the two methods.

Analysis of the pilot data revealed that ratings administered via the company's e-mail system had a higher mean, a narrower range, and more blanks than ratings taken from optically scanned paper forms. The distribution of the scores was also markedly different: Paper-based ratings were distributed along a normal bell curve, indicating reliable and valid results, while ratings from the company server were strongly skewed toward favorable answers. These results suggested that respondents were reluctant to provide anything other than unrealistically favorable ratings of their leader and peers when they knew that their responses were being compiled somewhere on the company mainframe. Duke now lets participants choose the format they prefer for the survey: a conventional paper form or a new Web-enabled version running on an external server owned by a third party.

15. In large organizations, make the department the primary unit of analysis for company surveys. While the need to retain anonymity is paramount, large corporations still need to organize and analyze the results of internal surveys at the department or operating unit level because they assess performance at those levels. Clearly, surveys that are undifferentiated by department will be limited in their usefulness. In designing large surveys, therefore, it is useful to add a check-off sheet (or a list of codes) identifying a respondent's facility and department. This feature helps you put together customized feedback reports that cluster departments and divisions into the precise groupings you need. Adding this feature to a large survey for International enabled us to deliver nearly 400 customized reports – some summarizing a single department's results, others summarizing sectors (a cluster of departments), facilities, or entire divisions – only one month after we collected surveys from more than 10,000 employees.

16. Make sure that employees can complete the survey in about 20 minutes. Employees are busy, and nobody really likes surveys and assessments. If a questionnaire appears excessively time-consuming, only people with a lot of time (hardly a representative sample) will participate, and the response rate will fall dramatically. We've already seen that when surveys are long, respondents' answers become automatic and overly positive. In general, we've found

Well-Designed Surveys Produce Normal Results

Well-designed surveys generate data that follow the normal bell curve: A small number of the results lie near the low end of the scale, most are average, and a few are exceptional. Poorly designed surveys generate skewed data that depict overly high or low responses.
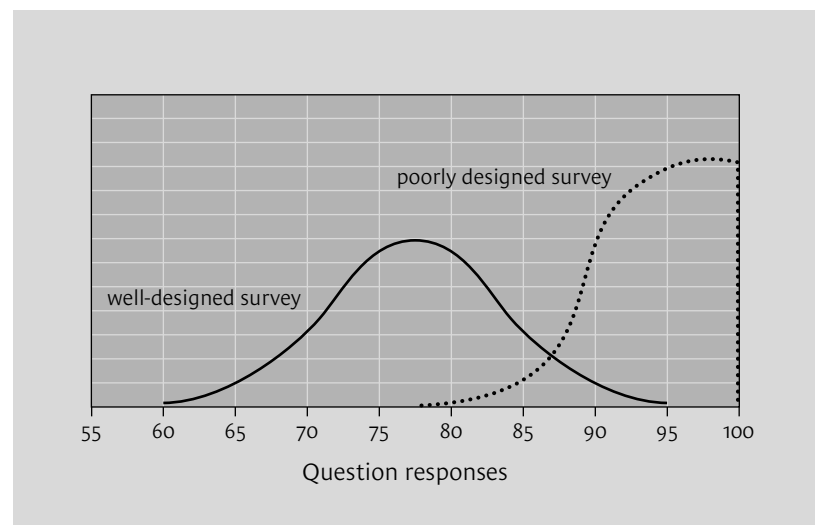
that surveys that can be finished in 20 minutes can provide substantial results for a company.

A sign at the auto parts store in my hometown states: "The wrong information will get you the wrong part … every time." Good surveys accurately home in on the problems the company wants information about. They are designed so that as many people as possible actually respond. And good survey design ensures that the spectrum of responses is unbiased. Following these guidelines will make it more likely that the information from your workplace survey will be unbiased, representative, and useful.