does the question influence the answer?

John M. Cavendish

If we were handed a tape measure and asked to measure a table, most of us could do that without too much trouble. Moreover, if a number of people were asked to repeat this task, the results of each measurement would likely be similar. If a yardstick were substituted for the tape measure, the task would still be simple enough, but there would be greater variation in results from person to person. Similar variance could be introduced by using an even smaller measuring device such as a ruler.

In the case just cited, we're talking about known measures and procedures subject to minimal interpretation errors. How desirable it would be if measuring human behavior could be done with such well-defined instruments and procedures. But, as most who are required to evaluate their programs will lament, nothing could be further from the truth.

The Theory

Program evaluation should measure changes that took place between Point A, a time preceding the program, and Point B, a time subsequent to the program. Of course, many variations on this simple design exist that allow us to have greater confidence that our results really measure program effects. However, this simple design shows some of the pitfalls awaiting those with program evaluation responsibilities.

The measured impact of a program derives from three main factors: the adequacy of theory on which the program is based, the adequacy of its implementation, and the adequacy of its measurement.[1] Educational programs,

John M. Cavendish: Extension Specialist, Health Education, Center for Extension and Continuing Education, Cooperative Extension Service, West Virginia University—Morgantown. Accepted for publication: May, 1983.

Extension's or others, certainly can be weak in any or all of these areas. However, measurement is the focus of this article.

How convenient (for those involved in measurement) if we lived in a dichotomous world. The answer to every question would be yes or no; black or white; effective or ineffective. Unfortunately, that's not the way the "real" world operates. A recent evaluation of a stress management workshop conducted for teachers serves only too well to show the difficulty encountered in measuring human attributes.

The Reality

The stress management workshop was designed to help teachers recognize the symptoms of stress associated with job burnout and to learn simple strategies to reduce this stress. A total of 55 elementary and secondary teachers, teachers aides, counselors, and administrative personnel participated. As a part of the evaluation instrument, three questions were designed to assess past and present job burnout. The following format was used.

Q1. Are you or have you ever burned out from job stress?
Yes—33% No—67%

Q2. On the following scale, indicate by circling a number where you feel you are:

0	1	2	3	4	5	6	7	8	9	10
7%	7	9	27	9	15	7	15	4	0	0%
Not burned out										Completely burned out

Q3. What stage of burnout do you feel you're in?
2% 1. Exciting job, no burnout symptoms.
63% 2. Occasional stress, no burnout, less energy.
31% 3. Burning out, one or more physical or emotional symptoms.
4% 4. Crisis stage, chronic symptoms that won't go away.
0% 5. Completely burned out, need help.

Question 1 was a yes/no type. There's a great deal of reluctance on the part of individuals to say yes to such a question, particularly if it involves a sensitive area.[2]

Question 2 was a partially anchored scale, anchored at the endpoints. A strikingly small percentage of the group (7%) indicated they weren't burned out at all.

Question 3 was a fully anchored scale, anchored at each point (stage) with specific suggestions about how one at that stage might feel or be. Only two percent of the group rated themselves completely free of burnout symptoms on this scale.

It was interesting to note that Q2 and Q3 showed a positive correlation to each other $R_s = .66$, while their relationship with Q1 was weaker. The correlation (R_s) between Q1 and Q2 was .38 and between Q1 and Q3 was .27. This suggests that answering "yes" to Q1 didn't necessarily mean one rated oneself higher on Q2 and Q3.

Implications

These examples indicate that the structure of questions on an evaluation tool can make a significant difference in the results of the evaluation. This study suggests the dichotomous (yes/no) type format has severe limitations when used in evaluation. Thus, to evaluate a stress management workshop by asking participants "Did you make any changes in the way you coped with stress as a result of the workshop?" would likely yield less flattering (and less accurate) results than using the other formats suggested. The problem with a limited choice format leads us to the question of how to select the best format for a given situation.

We should structure our questions so, that to the extent possible, a natural relationship exists between the response categories and the *behaviors or attributes* we're measuring.[3] Further, the response categories should reflect the spectrum of possible responses to the questions asked. For example, if we wanted to know how a group felt about spaghetti, asking the question "Do you like spaghetti? YES__NO__" would hardly give us the quality of information we could obtain by using the other types of ratings discussed. A better format would be to ask:

Compared to other foods, how do you like spaghetti?

0	1	2	3	4	5	6
Like it the least						Like it the most

Yes/no responses should be used only in instances where they cover all or practically all of the possibilities. A question such as "Is this the first Extension-sponsored program in which you have participated?" can appropriately use a yes/no response format, whereas it's not appropriate for a complex topic such as burnout.

A partially anchored scale, such as used in Q2, is useful when the exact categories of possible response aren't

known, but we do have some idea of the extreme responses. It's important that descriptors used for the endpoints actually represent the extremes. Descriptors such as *never* and *always* are better than seldom or frequently because they're less subjective and are easily seen as extremes. It's desirable to select descriptors whose meaning will vary least from individual to individual. However, it's a good idea to use zero to represent the negative endpoint since it can be easily recognized as the absence of the trait we happen to be measuring.

We should structure our questions so, that to the extent possible, a natural relationship exists between the response categories and the behaviors or attributes we're measuring. . . .

The number of intervals used in such a scale is usually five or seven. Ten, as the positive endpoint, may have some utility since we're used to thinking of the decimal system, although a great many researchers prefer to use five or seven. The partially anchored scale can prove very useful in measuring such subjective variables as to how valuable a program was perceived by participants, how effective the presenter was in conducting the program, or the likelihood that the participant will take some particular action as a result of the program.

The fully anchored scale, such as used in Q3, is most appropriate when the trait being measured is complex and we have sufficient information to describe the categories or stages at which people might find themselves. In the instance of burnout, results from numerous studies have identified characteristics of various stages in the burnout process. This scale was probably the best format for the burnout question, since it gave a somewhat ambiguous construct some concreteness using specific symptoms.

An example of an area in which the fully anchored scale would be particularly useful is weight control, where the end result (how much weight is lost) is a result of a rather complex set of behaviors (keeping records of food intake, activity logs, calorie counting and reduction, increased physical activity, etc). In this case, various degrees of compliance with the program can be anchored to a number of response categories.

Does how we ask the question determine the answer? Yes, the very act of assigning response categories determines the type of information we'll get back. The time we spend in constructing an evaluation tool may be the most important time we spend on an entire program.

Footnotes

1. Lawrence W. Green and others, "The School Health Curriculum Project: Its Theory, Practice, and Measurement Experience," *Health Education Quarterly,* VII (Spring, 1980), 14-34.
2. John M. Cavendish, *The Development of a Drug Use Consequences Checklist for a College Population* (Ann Arbor, Michigan: University Microfilms International, 1983).
3. Nigel Lemon, *Attitudes and Their Measurement* (New York: John Wiley and Sons, 1973), p. 72.