2

Issues in Designing the Standardized Questionnaire

Greg Mason

Many evaluators take the questionnaire for granted. In the usual evaluation framework, first evaluation issues are specified and they generate evaluation questions. In turn, these evaluation questions are associated with indicators. Evaluators often translate evaluation questions directly into a specific item on a questionnaire. Many times this translation is not considered and it appears to be assumed that respondents are clear on the meaning intended by the researcher. Evaluators also seem to assume that a single question will provide all information needed to support an indicator. The principle of multiple methods extends to using multiple questions to support a single indicator.

Questionnaires have a number of important features which makes them biased data collection tools. First, all questions are intrusive. They disturb the respondent and request information in contexts where there usually no compensation for the effort in providing an answer. Interviewer skill and question phrasing may mitigate the intrusion, but eventually all questions become wearying. Second, and more important is that questions usually change people's minds. Rarely have respondents thought about an issue or problem in the sequence or manner presented in the questionnaire. The standardized instrument, as opposed to the key informant interview, presents a strict order to thinking about a problem. The complex problems of *inter-item contamination* (questions which influence the responses to subsequent items[1]) is a manifestation of the general problem that a questionnaire changes the respondent. Third, *social desirability bias* (respondents change their answers to appear in a better light to the interviewer), poor memories and incomprehension are common and subvert the overall validity of evaluation research.

These threats are well-known. Despite them, most survey respondents willingly participate in the questionnaire process and genuinely seem to make the attempt to answer questions forthrightly. Provided the purpose of the evaluation is

---

1 The terms "questions" and "items" are used interchangeably in this chapter.

accepted by the respondent, the questions are grammatical, and the interviewer is trusted, the integrity of most questionnaires may be maintained.

The next section of this chapter reviews the basics of questionnaire design. These rules are often altered to meet unique situations. Each "rule" should be seen as an articulation of a principle and not a hard directive. The subsequent section reviews some of the recent literature on question ordering and phrasing. This work, which has accelerated in the past decade, illustrates some of the important issues which confront the evaluator in collecting opinion on complex issues, such as whether the program has had an incremental effect. The following section reviews problems in collecting data about two specific areas common to many evaluations: employment status and ethnicity. Social assistance and other income-tested human service programs frequently *must* collect information on employment status, but this requires many more questions than is common to include on a questionnaire. Ethnicity is often included as an explanation of program uptake, yet measurement of this concept is elusive. The last section concludes the paper with a review of some simple procedures to improve a questionnaire. The core theme in this paper is that an experimental approach controls bias in the standardized questionnaire.

BASIC PRINCIPLES OF QUESTIONNAIRE DESIGN

An Example: Perceptions of Television Violence

Ensuring the integrity of the researcher's intent to the respondent, as well as the clarity of the response, is the basic problem in questionnaire design and administration. Interviewer effects and other distortions which occur in the field are not discussed in this chapter. Here, the only concern is the question phrasing itself.

Belson (1984) reports on a series of experiments conducted to assess the integrity with which the researcher's meaning of a question matched the respondent's interpretation. One experiment involved a question asked as a *semantic differential*. This type of question design consists of a statement, to which respondents agree or disagree. When the agree/disagree is expanded to include a range of feeling (such as "strongly agree, agree, neutral, disagree") a Likert scale is used. The specific statement was *"Television shows are too violent for children."* to which respondents replied "agree" or "disagree" in a telephone interview. Belson constructed an experiment in which the respondents were re-interviewed a few days later and debriefed on the meaning of the question. When re-interviewed and asked what they thought the words "television shows" meant, respondents provided a wide range of interpretations from "all TV shows" to "prime time." When asked what they meant by "children" variability also emerged. Some were thinking about pre-

schoolers, others had anyone under 18 in mind. Respondents were apparently interpreting the question in a variety of ways which may or may not have been intended by the researcher.

These variations in interpretation reflect deep misunderstandings between the evaluator and the respondent. If there are three distinct interpretations of the concept "TV show" and two distinct interpretations of "children", then rather than one question, there are six. Three main problems now exist. First and most important, the researcher has no idea of who answered a particular question version without additional probing. Second, even if the varieties of question interpretations were understood and matched with respondents, the sample would be spread across these separate categories and reduce the statistical power of the analysis. Third, not all and possibly none of the interpretations may be valid within the context of the theory which the researcher is testing.

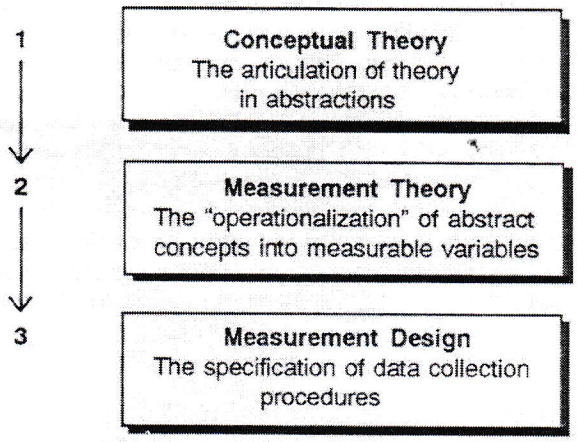
This example illustrates the most common defect in question phrasing. Evaluators may be aware of this problem and may simply accept the lower precision in the analysis. This bias is not readily apparent during the statistical analysis and is not detectable using formal methods. Because of this it is easy to simply accept the problem. This may reflect laziness or constrained research resources. Either way some basic principles of questionnaire construction assists in mitigating these errors.

BASIC PRINCIPLES OF QUESTION DESIGN

The problem of divergent interpretations on the part of evaluator and respondent reflects basic issues in question design. Underlying any program evaluation is a specific theory of how the intervention is intended to operate. Program theory in turn reflects a variety of political, social, and economic theories of human and institutional behaviour. To understand how to design better questions, the first step must be to articulate the theory underlying the program. This may be seen as a series of design steps in the early stages of the evaluation as shown in Figure 1.

Figure 1 shows the process of moving from theory to questionnaire design. The Conceptual Theory is the articulation of a program in terms of how interventions are believed to operate. For example, a program to encourage drug cessation among teenagers may use popular music stars to sell the message, seminars to provide support among peers, as well as active and visible drug enforcement. In this program the Conceptual Theory might articulate a model where role models are used to encourage appropriate behaviour. The assumptions about youth wishing to emulate their pop start idols and adhering to group norms are also part of the model. The extent to which the Conceptual Theory is able to articulate and detail the relationship between the theory of how these interventions operate and outcomes, determines the detail in the logic models and evaluation framework.

Figure 1.
The Process of Moving from Theory to Questionnaire Design



Many evaluators appear to not understand the importance of Measurement Theory which is an intermediate step. Conceptual Theory stage specifies program operation in terms of abstract relationships, while the Measurement Theory describes the variables which will be used to measure the theoretical concepts. For example, Conceptual Theory might use concepts such as "alienation", while the Measurement Theory might specify these concepts as attitudes toward the future and scope of independent action. The translation from theoretical concept to specific measure is a critical development in the evaluation design and framework.

Often this translation is presented as a single step, but it is more useful to consider it as two phases. The first phase specifies the *theoretical* measure, while the second phase translates it into precise question wording.

Theoretical measures or themes are often operationalized by several questions. Some variables are easy and unambiguous to pose and require only one measure. For example, much of the factual data on the respondent's demographic and economic attributes may be obtained with a single question. Age needs only one question, although there are various ways of asking this question, some more efficient and accurate than others. Other concepts may lead to several theoretical measures, each of which in turn may require specific items to capture. Formal

analysis (statistical tests) not only test relationships as reflected by the variables, but also evaluate the adequacy of the specific questions to support the theoretical measures.

There are two basic principles for question design. *Grounding* refers to locating the question within the everyday context of the respondent. *Bounding* refers to the imposition of limits to contain the range of interpretations. In the case of the example of TV shows and violence, bounding confines the frame of reference to specific times such as "evening television between 7:00 pm. and 10:00 pm." Further bounding refines "children" to "children under the age of 6."

Grounding addresses the phrase "too violent" by specifying specific acts. A potential refinement might be to use the term "encourage children to become violent." In this way a general value statement about too much violence, is converted to opinion about changes in behaviour. Grounding works on the focus of the question and requires care to ensure that the intent of the theoretical measure is preserved. Bounding determines the range of possible responses. Using a spatial analogy, grounding refers to the point in space, bounding refers to the area around that point.

COMMON "RULES" FOR DESIGNING ITEMS ON THE STANDARDIZED QUESTIONNAIRE

Over the years, a number of common rules for questionnaire design have become established. These rules, or more accurately principles, provide a checklist for the evaluator to examine each item in the survey.

1. Set the level of wording to the respondent. For the general public the median literacy level is Grade 8. For technical audiences, jargon is permissible as long it is certain that all respondents will understand the terms. As an example, the term "drug companies" is used for the general public, and for doctors "pharmaceutical companies" is more appropriate. Aiming too low results in loss of credibility among some audiences, while aiming too high may confuse and discourage the general respondent.

2. Use short questions. Some issues are inherently complex. If the respondent requires information to make an intelligent response it is preferable to use several questions to "set the stage" rather than trying for a lengthy question. This may raise problems in question order and interaction which are discussed in the next section. Long preambles test the patience and skill of interviewers, since not all will read the test with equal clarity and rhythm. Long texts tire the respondent as well.

3. Balance alternatives. It is always better to state there are two (or more perspectives), then to cite the alternatives, and finally to ask the respondent to chose. To simply ask "Do you think the Apply Valley irrigation program is worthwhile" contains possible bias. A superior alternative is a dichotomy such as "Some people think that the Apple Valley irrigation project is worthwhile, others think it is not, what do you think?" Care must be taken with this type of question since the order of the alternatives may "load" the question one way or the other. Also, it is easy to subtly bias responses in the description of the alternatives. There is not much danger in flagrant bias, since many respondents will object to overt manipulation. Rather, subtle, unconscious shades of meaning are the most common forms of distortion. Pre-testing and the use of many readers/listeners is essential to reducing this threat to validity.

Question loading can take a number of forms. Stating a false premise *"In order to balance the budget, should the government reduce expenditures on roads"* assumes the respondent believes the budget should be balanced. Even if interviewers state *"Well assuming you agreed with balancing the budget ..."* the bias remains.

4. Focus the question to obtain unambiguous answers. Using response categories which are not mutually exclusive encourages vague answers. *"Do you sometimes go to the Smith Street drop in centre, or do you sometimes go to the one on Main Avenue"* is a typical example. Another example of poor focus is the *double-barrelled* question. *"Are you happy with the range of courses and hours of operation of the downtown employment training centre"* is an example of this bias. There are four general responses to this question (satisfaction with both, satisfaction with one and not the other, dissatisfaction with both), but there are only two response categories allowed.

5. Avoid hypothetical questions. Most people are poor predictors of their own behaviour. Needs analysis which ask individuals to predict future requirements, or pose a hypothetical service and then ask whether it will be used usually generate very poor information. "If a teen drop-in centre were opened in Smith Street, would you go there" will typically produce an overestimate by a factor of at least 2. Past behaviour and the present situation are superior indicators than promises. Someone who never has been to a drop-in centre is unlikely to start unless circumstances in his or her life have materially changed.

6. "Don't Know", "Neutral", and "No opinion" are different. In the interests of increasing the appearance of validity, researchers sometimes suppress the critical differences among these three concepts. "Don't know" should refer to the respondent who has not considered the issue and cannot make any judgment, "No opinion" should refer to the respondent who has considered the issue and is disinterested in any alternative, while "Neutral" should refer to the respondent who

has considered the issue and come to a position between the extremes. Separating these concepts is tricky. First, most respondents will be reluctant to state they have not considered important social issues. Second, ignorance and indifference are unacceptable and will not be revealed.

Potential resolutions include phrases such as " ... *many people we have been talking to say they do not have enough information to make up their mind. If you feel the same way please tell me.*" Including the middle position in listed (or read) Likert response categories is also useful for encouraging the "Neutral" response.[2] The middle position is often suppressed if the range of opinion and attitude is important. Respondents must be encouraged to believe that responses which indicate lack of knowledge or indifference are common and acceptable. Evaluators should design questions to obtain maximum information and discriminating among the "Don't Knows", "Neutrals", and "No Opinions" may have important policy implications. For example, those who don't care or who have insufficient information may require educating while those who have knowledge and are neutral may represent a centre of gravity of opinion. They may also have partial knowledge and could swing one way or the other after additional information. A critical aspect of item design is to allow respondent to express their full opinion, either by probing on a verbal interview or by encouraging verbatim comments to be written on a mail survey.

Open and Closed Questions

Expediency often dictates that questions be closed. A closed question presents a fixed listing of response categories where there are no opportunities for the respondent to record an answer which is not on the list. An open question allows a complete verbatim transcript of the respondent's view. Intermediate approaches where the category *"Other (please specify)"* attempt to include verbatim information but these should be treated as closed, since most researchers will rarely examine the actual phrases or words used, unless the percentage in this category is high relative to other responses.

Technically the open question is superior since it always provides more information than fitting the respondent into closed categories. But the need to process high sample numbers to obtain statistical validity usually eliminates the verbatim responses. Often a question may be open at the start, but as response categories are identified within the recorded information, questions are closed to increase the speed of survey administration. Budget constraints are a reality and it is inevitable that standardized closed item questionnaires will dominate evaluation research.

---

[2] Compare the scale "Agree or Disagree?" with "Agree, Neutral, or Disagree?" and with "Agree or Disagree, or have not formed an opinion yet?"

QUESTIONNAIRE DESIGN

Sensitive Issues

Personal information and family secrets are difficult to collect. Income is a prime example of a question many people find sensitive. Simply asking what a person's annual income is does not work. Many people do not have a good idea of the their annual income. Self-employed, entrepreneurs, farmers, and sales persons are examples of people who may have a poor idea of their annual income.

A second problem is that sensitive issues come "too close." A request for exact income is much more threatening than asking the respondent to supply their income within a broad range. Respondents are much more comfortable with a question that asks *"I am going to read a number of broad income categories. When I come to the one which applies to you, please stop me."* The broad ranges ($10,000 increments) allow respondents to reduce the detail pertain to them, and stopping an interviewer reciting ranges is not the same as divulging confidential information. Similarly, asking someone their year of birth, is much less threatening than asking him or her how old they are.

Some issues are very personal and touch deep problems within the family. Studies of victims often must probe into high personal areas. The standardized interview is usually a poor method for this task. In some cases an evaluation must seek information from the general public on awareness or experience with severe social problems. For example, asking the extent personal experience with child abuse may pose difficulties for those who experienced the problem, or who are currently dealing that problem within their family. A direct probe will often be met with complete denial and can result in abrupt termination of the interview. Indirect probing may be successful. Rather than asking *"Does your immediate family experience ...."* it is better to move obliquely and ask *"How serious is .... within your immediate neighbourhood?"* Although it is difficult to validate, those who say the problem is very serious probably have first-hand experience.

The standardized interview is typically a poor approach to dealing with personal and sensitive issues. Discussions and less structured personal interviews are typically superior. However, using an indirect approach and allows the respondent to distance themselves from the threatening aspects of these types of questions produces success.

The Middle Position

There is debate over the middle position and whether or not it should be retained as part of the response alternatives. One school states that it allows respondents to evade making a choice, while another argues that the middle position is a valid response and should be encouraged. Schuman and Presser (1981) evaluated a number of alternatives and used "split ballots" to examine response

effects to questions with and without the middle position. In the end they concur with Payne's (1951) classic directive:

> *If the direction in which people are leaning on the issue is the type of information wanted, it is better not to suggest the middle position .... If it is desired to sort out those with more definite convictions, then it is better to suggest the middle position.*

If it is possible to ask several questions, then specific questions probing for tendencies is probably preferable to forcing the issue with a single question that omits the middle position. Respondents may resent not being provided with that middle option, but they may offer a tendency after a few exploratory questions. Interviewers also find it stressful to deal with respondents who provide a middle response to a question which does not include that category. Either it must be coded, or the interviewer must re-ask the question and request a choice. In both situations the objective of economy by using a single question is compromised.

Summary of Basic Principles

The basic principles of questionnaire design are clear. By grounding and bounding questions, respondents are channelled into specifics. This reduces bias and provides a firmer foundation for program evaluation and design. Recent advances in questionnaire design illustrate that these principles support a wide variety of formats. Contrary to being a rigid model of questionnaire design, the principles allow evaluators considerable latitude in communicating effectively with respondents. The key objective is always to ensure that the respondent understands the question as intended.

QUESTION ORDERING AND RESPONSE EFFECTS

There is a growing literature on question ordering. Awareness has emerged that the questionnaire is never neutral. Until the time of the interview most respondents may have thought about the subject matter only casually, if at all. The standardized interview places the typical respondent into an intensive process of "structured consideration." This is unusual and alien. Further, the order of the questions is important. Each question forms background to which other questions relate. This may be exploited beneficially as questions provide a context and allow

complex ideas to be introduced gradually. A negative effect occurs when questions lead to a conclusion, or when one question has a significant impact on the responses provided to subsequent issues.

Start Simple and Keep it Interesting

The first questions must be simple and non-threatening. Many respondents disqualify themselves under the belief that only "experts" have valid opinions. For example on a recent survey undertaken by the author, the introduction mentioned that the questionnaire dealt with nuclear war and Canada's role in strategic defence. Many respondents, especially older women stated that they did not know anything about the topic and refused to participate. (Men typically are less inclined to admit lack of knowledge on technical and military matters). The introduction was changed to " a discussion of war and peace." Refusals dropped dramatically, possibly because some thought the topic was dealing with a great novel, but more likely that the topic non-threatening and non-technical.

Another example illustrates the importance of using the first questions to build respondent interest and commitment to the questionnaire. On a recent energy conservation study by the author, a series of policy evaluation questions were included to create interest in an otherwise dull questionnaire dealing with the extent of insulation in the home. These policy questions concentrated on respondent assessment of government policy, and were significant in raising response rate.

Factual and Demographic Questions at the End

Many questionnaires start with a dirge of demographic questions. Despite the ease with which respondents can answer these questions it is a boring way to start any questionnaire. Also, some demographic questions, e.g., income are inherently sensitive and can cause early termination of the interview.

Demographic questions should be left until the end the interview. If the income question does cause higher refusal rates, its omission is less problematic if other data have been collected. Also, after a relationship with the interviewers has been developed, it is less likely that sensitive questions will be refused. The need for these data should be rationalized to the respondent as "to ensure we have a statistically valid sample of Winnipeg." Most respondents accept this as reasonable.

Consistency

It used to be common to include questions to test for consistency of response. The basic idea seemed to be that a question repeated with a slightly different phrasing, or in the negative[3], should be answered consistently. This practice is quite dubious.

First, if inconsistency is detected in a standardized interview, there may or may not be any opportunity to clarify the respondent's "true" position. It seems an unlikely proposition to recontact respondents to re-ask questions — some might take offense at the suggestion they were inconsistent. This approach is also expensive. A major advantage of the in-person interview is that an alert interviewer can probe for clarification; when the respondent has an opportunity to explain a position in more detail inconsistency sometimes evaporates. This careful probing is usually not possible with mail surveys and is less feasible for a telephone compared to in-person questionnaire.

Second, it is well to remember that inconsistency is simply part of any individual's opinion. Most of us do not have a completely seamless value system, and we commonly discover that part of our opinion, when carried to a logical conclusion, may turn out to be inconsistent with some other beliefs.

Evidence of inconsistency must be very carefully evaluated. A directly related question, such as repeating a statement in the negative to test for a response reversal, must be distinguished from a shift in response as a result of a phrasing variation. The variation may be sufficient to be interpretable as a different question.

Third, if inconsistency appears obvious, what is the appropriate course of action? In general, inconsistency should be interpreted as a failure in question design and technique. Pre-testing and initial reflection may have revealed that respondents had conflicting views on a particular program or policy. In addition, there may be untapped aspects of the program and its delivery — inconsistency could be an important indicator of these issues.

Carry-over Effects

The general problem of question order may be described by the effect that pre-ceding items have an influence on subsequent responses. To reiterate the introduction it is very important to remember that until the interview starts, most respondents may have not thought about the issues. Even raising certain topics may influence responses to items latter in the interview.

---

[3] The semantic differential format lends itself to repeating statements in the negative, to which the response should be reversed if there is to be consistency.

"filtering" refers to actions taking by researchers to reduce the distortions produced by these pseudo-opinions. First, questions are used to increase the awareness level of the respondent. In addition, questions can contain explanations of fact. *"As you may know, the social services department has declared a deficit for the third year..."* The filtering questions and statements have serious potential for introducing bias and restraint must be exercised.

The second sense of filtering is to encourage respondents to reveal ignorance or that they have not thought about the issue. Adding statements such as *"Many people say they are unaware of this so if you do not feel you have enough information to make a judgment please tell me."* There is sometimes great pressure to submerge those with weak preferences or opinions. In political attitude surveys, a high number of respondents who state they are undecided is sometimes interpreted as resulting from poor questionnaire design. However, knowing how many people really do not know, or who have weak opinions is intrinsically important for many programs.

Inter-Item Contamination

The interaction of items on a survey is well known. A general question on how well the government is doing in health policy, will often produce different responses if it leads or follows specific questions on major problems such as AIDS or cancer. One approach is to place the general policy question first, and then follow with questions on specific problems. A general health policy question is followed by items dealing with aspects of the health care system. Proponents of this approach argue that this insulates the general questions from contamination by the specific items. The other view is that general questions should follow specific items. In this way, the respondent is encouraged to reflect on specifics of an issue before rendering a general judgment. This latter view is in line with the above discussion on filtering and is the preferred approach. It is very important that the specific items on aspects of a program or policy are unbiased and complete. The order in which these specific items are presented may also be important.

Resolution of Order and Response Effects

Question ordering and inter-item effects are commonly understood to be serious problems, but there is little progress in developing general design guidelines. This should not be surprising since each questionnaire is unique and generates its own set of problems. It is more fruitful to approach questionnaire design from an experimental perspective, where these contaminating effects are controlled through randomization.

For example, it is well understood that bias results when a list is read and respondents are requested to make a selection. On short lists the first response alternative tends to be selected, while on long lists the last mentioned item tends to be selected more frequently. Randomizing the list order is used to control this effect.

The concept of randomization can be extended from response lists to question order. The *split ballot* involves dividing the sample into two or more groups. Questionnaires are identical for each group, except for specific changes which are thought to be subject to order effects. Respondents are randomly allocated to various groups. Statistical analysis is used to estimate the impact of question order and wording by testing whether response patterns differ among the various groups.

The use of split ballots is relatively infrequent considering the pervasive effect of inter-item contamination and uncertainty over wording. If significant differences are found among the questionnaires[4], the evaluator must then examine the questions in more detail. There are many pressures to avoid an experimental approach to design. Often the client finds it confusing to have more than one version of the questionnaire. Evaluation budgets are usually strained and split ballots are more expensive because samples have to be increased to allow each questionnaire version to have sufficient information for statistical analysis. Also, the analysis itself is more expensive.

QUESTION PHRASING
THE ROLE OF SUBSTANTIVE KNOWLEDGE

As research has proceeded in substantive areas (as opposed to methodology), insight has grown to support more informed questionnaire construction. Two examples serve to illustrate how substantive knowledge improves questionnaire design.

Measuring Joblessness

Many evaluations of social services and income maintenance programs need to measure unemployment and employment. The general question, *"Are you employed full-time or part-time?"* appears clear and straightforward, however many biases confound this concept. To understand the problem that many people have in reporting their labour force status, recall the definition of the labour force. The labour force consists of adults over a particular age who are employed for pay, or who are

---

[4] By significant we mean a difference in response patterns which is greater than the margin of error for the samples involved.

looking for work. Unemployment consists of those members of the labour force who either cannot find any work, or who are unable to secure sufficient work to maintain the level of income desired.

The key areas of ambiguity relate to the extent to which an individual is looking for work, and the extend to which an individual obtains less work than desired. For some, job search is a full-time activity and may consist of many hours in the day devoted to resume preparation, calling on prospective employers and following up on leads. Others pursue a much more causal approach to job search. The discouraged worker phenomenon is also well known and describes a group of potential workers who, for many reasons, have decided that looking for a job is not worthwhile.

Another important phenomenon of joblessness relates to workers who are very selective in their job search. Highly skilled workers will often elect to remain out of work in the expectation that an opening in their trade might occur in the near future. Their skill level commands a higher wage, and it makes economic sense to turn down low paying jobs, if there is a reasonable expectation that a higher paying position is imminent.

Time is also important. Unemployment may be measured at a point in time or over span. This perspective views working as a flow of worker hours, not a state of affairs. Unemployment is lost labour time. Official statistics use the previous week as a reference period. This is more accurate than simply collecting information on whether the respondent is currently unemployed.

Clarifying the extent and nature of job search is obviously important. For this reason, most official measures of unemployment take pains to ask questions to determine whether the respondent was actively seeking work during the reference period. Other questions may be used to identifying those who had become discouraged after prolonged fruitless searching. Without these qualifications, reported unemployment is usually at considerable variance from reality. Most respondents are reluctant to admit their jobless state and this social "fact" is under-reported. When general unemployment rises being without a job is not unusual and the reluctance to disclose will fall.

Evaluators who ask a single question on employment status should understand that social desirability bias will confound this question. If employment status is important to assessing program effect, then the questionnaire should spend more time (ask more questions) to properly isolate this phenomena.

Ethnicity

Ethnicity is often identified as a "key" variable in social research. For example, the income maintenance experiments of the sixties and seventies found that ethnic

background appeared to be associated with willingness to work at various levels of income support. Ethnicity is often used to explain anomalies in evaluation results and many evaluations routinely include a question on ethnicity.

The concept "ethnicity" is complex and subtle. It has at least four distinct meanings - race, nationality, religion, and language spoken. Furthermore, ethnic identification may change. Upon immigrating to a new country, some may identify strongly with the new land and state they are "Canadians" while others will retain a strong attachment to their origins. Children of immigrants may lose their ethnic identification when they start school, often to the consternation of their parents. Later, as they mature, the second generation, may recover previous ethnic roots and state they are "Greek Canadians."

Ethnic identification is sometimes in the mind of the respondent. Race is not, and neither is country of origin. Evaluators who wish to remove the subjectivity of a question such as *"What do you consider to be your ethnic affiliation?"* have several choices. One is to focus on language and a common question is *"What language did you first learn as a child?"* A stronger language question is to ask whether the respondent still speaks that language regularly. Another tack is to probe for country of origin. More subtle questions probe for the heritage of parents and grand parents.

Race is more a difficult and charged subject. Sometimes, this are included in administrative data, but the number of categories is often constrained. Also, the categories can be loaded and difficult for many to ask and answer. For example, "White" and "Caucasian" tend to be treated as synonymous, while "Black" and "Negro" are not.

The typical ethnicity question may combine all these categories. A questionnaire may ask generally "What is your ethnic affiliation" and then include response categories such as "Black", "French", and "Jewish" on the same list. This mixed response list produces total confusion because the categories are not mutually exclusive. It is mandatory to decide which aspect of ethnicity is relevant for the program and then probe along that dimension. If more than one aspect is important, use separate questions. If all four aspects of ethnicity are important, a very complex classification scheme may emerge. This can produce undue complexity in the analysis and probably should be avoided. For many purposes the langauge first learned and country of origin are the most useful concepts to classify ethnic background.

SUMMARY

There is no such thing as a single question on ethnicity or unemployment, yet the vast majority of surveys are designed as if this was the case. These are complex concepts, but if the questionnaire is to be managed, choices must be made. Decide on the relevant aspect of a concept, and then explore it.

From the above discussion it is possible to identify a number of basic principles for questionnaire design:

- Literature searches provide the substantive knowledge to design appropriate questions. Further a proper literature review may well produce previous surveys which can be used as a basis for an evaluation. Most valuable is material which critically evaluates individual items in a survey.

- An experimental approach to deal with potential inter-item contamination and order effects. Rather than a rarity, split ballots should be quite common. Randomization of response lists and testing question versions with randomly selected subsets of the main sample will control many of the otherwise undetected biases in the data.

- Separate the questionnaire "pre-test" from the "field test." The pre-test ensures that respondents understand the question as intended. This involves administering the questionnaire and then debriefing respondents to determine what is understood. Follow-up telephone calls, in-person interviews and focus groups are all useful ways to determine how respondents interpret a question. On the other hand, the field test ensures that interviewers can administer the questionnaire, that response rates on mail surveys will be adequate and that the mechanics and logistics of the survey are properly organized.

- Survey research continues to evolve. Increasing complexity of question design is likely, and this means that the randomization of the split ballot will also continue to be developed. Experimentation and the resulting statistical sophistication will grow as an inevitable part of sound methodology.

FURTHER READING

This set of references will provide an initial taste of current issues and research. Use them to track down other references as your interest dictates.

Belson, A., (1984). *The Design and Understanding of Survey Questions*, Aldershot, U.K.: Gower Publishing.

Bishop, G.F., Oldendick, R., Tuchfarber, A., & Bennett, S. (1980). Pseudo opinion on public affairs", *Public Opinion Quarterly*.

Bishop, G.F., Oldendick, R., Tuchfarber, A. (1985). The importance of replicating a failure to replicate: Order effects on abortion issues", *Public Opinion Quarterly*.

Kalton, G., Collins, M., & Brook, L. (1978). Experiments in wording opinion surveys, *Applied Statistics*.

Labaw, P. L. (1981). *Advanced questionnaire design*. Cambridge, MA: Abt.

Payne, S. (1951). *The art of asking questions*. Princeton, NJ: Princeton University Press.

Perrault, W. (1976). Controlling order-effect bias, *Public Opinion Quarterly*.

Presser S. & Schuman, H. (1980). The measurement of the middle position in attitude surveys, *Public Opinion Quarterly*.

Schuman H. & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.

Tourangeau, R., Rasinski, K., Bradburn, N., & D'Andrade, R. (1989). Carryover effects in attitude surveys, *Public Opinion Quarterly*.

Greg Mason, Ph.D. is a Principal of Prairie Research Associates, Inc. in Winnipeg, Manitoba. He is also with the Department of Economics at the University of Manitoba.