The Canadian Journal of Program Evaluation Vol. 11 No. 1 Pages 73–94
ISSN 0834-1516

RECENT ADVANCES IN QUESTIONNAIRE DESIGN FOR PROGRAM EVALUATION

Greg Mason
Prairie Research Associates Inc. and
Department of Economics, University of Manitoba
Winnipeg, Manitoba

Abstract: Evaluation uses questionnaires as a central data-gathering technique, yet researchers often appear unaware of recent developments in questionnaire design. This article reviews issues beyond the creation of standardized questions and the basic rules researchers find useful in data collection. These elementary guidelines remain robust for much evaluation research and should not be abandoned hastily. However, rapid change in the theory underlying questionnaire design has important implications for evaluation. Three themes illustrate these changes. First, magnitude scales and their use in client satisfaction scales show how response categories can improve individual questions. Second, decision theory sees respondents as selecting "correct" responses from a portfolio of potential answers. In this view, the answer to a question is conditioned by the values held by the respondent and his or her perception about the risks of revealing true feelings. Third, in some cases it is possible to cast the entire questionnaire into a framework that replicates how choices are made by respondents. In one application, the questionnaire simulates the decision-making of a consumer in the market place. Each of these three themes is applied to problems evaluators face in data collection involving surveys and questionnaires.

Résumé: Même si l'évaluation a recours au questionnaire comme technique fondamentale de collecte de données, les chercheurs semblent souvent ignorer les progrès récents de la conception des questionnaires. Dans le présent article, on examine des éléments qui vont au-delà de la création de questions normalisées et des règles de base que les chercheurs trouvent utiles pour la collecte de données. De telles lignes directrices demeurent valables pour la plupart des recherches dans le secteur de l'évaluation des programmes, et elles ne devraient pas être abandonnées à la hâte. Par contre, l'évolution rapide de la théorie sous-jacente à la conception des questionnaires a des incidences

importantes sur l'évaluation. Trois thèmes illustrent les changements. Premièrement, les échelles de grandeur et leur utilisation dans l'établissement des échelles de satisfaction de la clientèle indiquent comment des catégories de réponse peuvent améliorer des questions individuelles. Deuxièmement, dans la théorie de la décision, on considère que les répondants choisissent les réponses «justes» dans un ensemble de réponses potentielles. Selon un tel point de vue, toute réponse est conditionnée par les valeurs du répondant et par sa perception des risques liés à la divulgation de ses sentiments réels. Troisièmement, il est possible dans certains cas d'élaborer le questionnaire selon un cadre qui reproduit le processus de choix du répondant. Le questionnaire, par exemple, peut simuler la prise de décision d'un consommateur dans le marché. Chacun des trois thèmes est appliqué à des problèmes auxquels doivent faire face les évaluateurs dans la collecte de données au moyen d'enquêtes et de questionnaires.

Many references explain how to ask good questions, and researchers often assume that asking a clear question is not that difficult. Payne (1951), Belson (1981), Sheatsley (1983), and Converse and Presser (1986) are commonly cited. Most sources on questionnaire design explore issues of syntax and meaning to derive rules for proper question phrasing and construction. These rules can be varied as appropriate for self-administered (mail), telephone, and in-person surveys.

Questionnaires seem to be developed according to a "vacuum cleaner theory" of data gathering. Many researchers see questionnaires as intended to "suck" maximum information from the minds of compliant respondents. We assume that every question is straightforward, that respondents are passive conveyors of data, that questions do not affect each other, and that fatigue becomes an issue only on very long questionnaires.

This review article examines recent developments in questionnaire design. No new ground is forged. We examine three ideas from this rapidly evolving area. First, researchers have tried to improve the clarity of the individual question and its response categories. As an example, we examine magnitude scales and their use in client satisfaction surveys that have become common in evaluations of public services.

Second, we review how respondents "frame" their answers and show how this is critical to understanding responses. Developments in

decision theory have changed the way researchers perceive that respondents choose their answers. Respondents are now seen as editing the information they choose to share with the interviewer.

A third development places questions within a framework that tries to replicate the context in which respondents make choices. Here the emphasis is on the structure of the entire questionnaire. Market researchers often attempt to place the respondent in a decision-making context that approximates the process of choosing from a number of alternatives. Techniques such as contingent valuation and conjoint analysis are examples of this approach and are discussed later in this paper.

Before reviewing these three themes, it is useful to touch on the difficulties that confront the questionnaire designer.

WHY IS QUESTIONNAIRE CONSTRUCTION SO DIFFICULT?

Whether in a conversation or a structured interview, the usual approach to soliciting information is to ask a question. Conversations combine questions, answers, unsolicited opinion, and reactions to what is said. They meander as participants question and probe. Often in a conversation we inadvertently show misunderstanding, and much of our time is occupied in clarifying confusion.

The standardized questionnaire is direct, time-limited, and stylized as it proceeds through a more-or-less-carefully developed sequence of questions. The focus is to ensure that the researcher's meaning is understood in the same way by every respondent. Often, we underrate both the difficultly in phrasing specific questions as well as the influence question positioning has within the questionnaire. Researchers tend to gloss over the differences between complex, multidimensional human conversations and the simplified, sequential process of a standardized questionnaire.[1]

The formal setting of the standardized survey interview imposes strict controls on what the interviewer may say to help the respondent, and on the range of responses allowed. Of course, in mailed questionnaires, the interaction between researcher and respondent is eliminated. This formalism appears to increase the researcher's control over the range of responses that may be provided, but as Belson (1981) shows, these assumptions are quite wrong. Respondents freely interpret question meaning. Based on examples cited by

Belson, we now know it is essential to leave no term unexplained or undefined. In the question "Are TV shows too violent for children?", we need to define "children" in terms of age and "TV shows" in terms of time of day. The researcher is challenged to define what "too violent" means. The need to clarify meaning, commonly called "grounding the question," stands in contrast to the usual dictum that questions should be concise.

The reference point of the respondent is also widely understood to be important. Culture and gender are frequent sources of variation in understanding between researcher and respondent. Men and women often hear the same question in quite different ways. Pretesting instruments, grounding the question and providing explanations for each term can reduce the variation in understanding. However, the cultural, social and economic class of the respondent influences interpretation and comfort in responding to a specific question.

For example, ratings of customer service are notorious in their "inter-rater variability" with some respondents providing a high score and others a low score to what is objectively the same level of service. As Devlin, Dong and Brown (1993) show, expectations condition perceived level of service. The reference point of the respondent must be understood before meaning can be extracted from the satisfaction scale. This may seem an obvious point, yet to clarify meaning, questionnaires must become longer, explaining context and not collecting the "real" information! This is difficult in an era of constrained evaluation resources.

IMPROVING THE INDIVIDUAL QUESTION

The first development in questionnaire design we discuss is at the individual item level. The typical closed-response category consists of categories with words that label discrete levels. However, words are slippery. Compare two common scales used in client satisfaction questions:

- Very Poor, Poor, Average, Good, Very Good
- Very Bad, Bad, Fair, Good, Very Good.

Although they appear superficially similar, a considerable difference exists in these two scales, especially in comparing the words "average" and "fair." Some interpret fair as less than average; others

think of fair as meaning something good or just. In the typical survey, respondents are allowed to interpret the response alternatives. Individual understanding is unknown to the interviewer unless a question is asked to clarify. Modern advertising has also rendered many adjectives without meaning ("stupendous, gargantuan, monster discounts," "Bill and Ted's very excellent vacation," etc.).

One can improve the Likert scaling process, especially in a telephone interview, by using a two-step process.

1. *What do you think of the program? Is it good or bad?*
2. PROBE BASED ON RESPONSE: *Would you say it is just good or is it very good (just bad or is it very bad)?*

This two-step question asks the respondent to provide a general impression, positive or negative, and then refine this degree of feeling. The five-point scale forces the respondents to "visualize" the scale and then position themselves within that spectrum. This is a mentally demanding task and the common use of seven-point scales in telephone surveys is difficult to understand.

The two-step version of the Likert scale presented above omits the middle position. Some researchers force the issue and do not accept the middle response. Recent work in client satisfaction (Waddell, 1995) shows that the middle position is critical in identifying the full spectrum of attitudes. Interestingly, in light of recent research into customer satisfaction (discussed below), Payne (1951) seems to have had the role of the middle position right. It highlights respondents with extreme views and allows management to determine what proportion of clients are truly pleased, as opposed to being just satisfied.

Despite the improvement produced by a two-step question, weaknesses remain in the conventional, word-based semantic differential scales such as those above.

- It is a truism in social research that measurement should occur at the lowest level. For example, a measure of automobile fuel economy can be ordinal ("good," "better," and "best"), interval (8 of 10), or ratio (100 kilometres per 10 litres). Compared to the ordinal scale used by most client satisfaction or opinion surveys, a ratio measure relies on physical measures.

- Reliability is compromised because the semantic space between words is unclear. What is the distance between "good" and "very good"? Is it the same as between "very bad" and "bad"? Lodge (1981) has good examples of the non-linearity between words commonly used in semantic differential scales such as very good, good, average, bad and very bad.

- The limited positions in the semantic Likert scale encourage positional and order response effects. With only "good" or "bad" as choices, respondents may change their rating if they feel forced in a specific direction. Respondents may give lower ratings for some aspects of the service to compensate for higher ratings awarded to other aspects.

Much of the current research into question construction seeks to refine the individual question as a data-gathering tool. *Magnitude scaling* focuses on creating a response scale based on a physical or numerical analogue so different respondents share a common understanding of the question intent.

To understand how this works, it is worth outlining what a good response scale should do:

- Respondents need to accept that subjective feelings can be translated to a semantic, numerical, or physical analogue.

- Different respondents should interpret the scale in same way. The response "good" should reflect the same subjective valuation for all respondents.

- A scale must discriminate among levels of perceptions. Respondents who rate a service as "fair" should be seen as reflecting a different and objectively (lower) level of service than those who rate it as good.

- Scale values should become a basis for action. A 7 out of 10 on a satisfaction scale should support constructive action for program management.

The magnitude scales as discussed by Lodge (1981) use physical devices such as a control for sound volume or the brightness of a light. Another device is a pressure gauge where the respondent grasps a ball and applies pressure. In each case the physical measure in terms of decibels, lumens, or kilograms of force is a gauge of

the degree of feeling. The respondent is asked to turn the dial to reflect their strongest feeling. This calibrates the physical scale, and all subsequent questions are normed to this maximum by having the respondent adjust the dial to reflect the subjective feeling associated with a particular question. This measurement approach is quite common. TV ads are evaluated using large audiences who turn dials to reflect positive or negative feeling about what is being shown at fifteen-second time intervals.

Social researchers have adapted this physical model to numerical scales:

> *Example: (telephone or mail)*
> On a scale of 0 to 10, where 0 is poor and 10 is excellent, what is your overall opinion of the program?
> 0 1 2 3 4 5 6 7 8 9 10

> *Example: (mail)*
> Using the scale below, please rate your last experience with the emergency room nurses.
> -2 -1 0 1 2
> (Very Negative) (Very Positive)

The first scale can be used in either a telephone or printed questionnaire, while the second scale, difficult for most to "visualize" on a phone survey, is best presented in a mail survey.

Magnitude scales are easy to understand if respondents are comfortable with translating subjective phenomena into a number. They are monotonic and linear, and they identify a middle position.[2] These scales also support quantification and summary statistics (means, variances, etc.), a feature that has made them attractive in client satisfaction measures.[3] This leads to calculating confidence intervals around the mean and awarding bonuses to managers if they improve satisfaction scores by a statistically significant amount.

Although magnitude scales are gaining in popularity and are an improvement on the Likert scale, difficulties can arise. First, and most obviously, respondents may not be able to translate feelings and beliefs into numbers. Worse, they may resent this format. However, respondents can still balk at expressing a complex feeling as a single number. In this sense the magnitude scale does not overcome the problems with the semantic Likert response categories.

Second, linear scales may not be appropriate for measuring subjective feelings. It could well be that the difference between "very good" and "excellent" is greater than between "good" and "very good" in the minds of respondents. Does the forced linearization distort the true intent of the respondent? Research is still clearly needed in this area.

Third, they tend to be overused. A battery of magnitude scales can cause confusion and respondent inattention as much as a long list of Likert scales. Restraint is needed when using this approach.

Moreover, recent research in client satisfaction is a little disturbing. Devlin et al. (1993) examine several approaches to client satisfaction measurement. They compare four general types of scales:

- satisfaction scales such as two-, four-, and five-point scales (e.g., very satisfied, satisfied, neither, dissatisfied, very dissatisfied);
- performance scales (excellent, good, fair, poor);
- gap scales in which performance is measured in relation to expectations (exceeded, met, nearly met, missed);
- non-anchored scales (e.g., a numerical scale from 0 to 10).

In their research, they find that many satisfaction scales do not discriminate among high and very high performance. These scales can fail to discover those who are satisfied, but have a specific complaint that has not been addressed. This can lead to growing dissatisfaction that could be uncovered earlier if the scale was more discriminating. The performance scale is universal in user satisfaction studies but, as with all semantic scales, it must be calibrated to the respondent. One way to do this is to ask the respondent to benchmark the service in relation to the best they receive. Citing comparable services is useful. For example, one can ask the respondent to compare a restaurant service with the best they have received in the last six months. This approach can become strained when one asks respondents to compare social service programs with the service provided by a bank. The benchmarking needs to encompass comparable activities.

In some cases, "massing" at the middle position or some other point can be obvious. For example, users are often reluctant to award a perfect "10" and many settle on providing a good pass such as a 7 or 8 out of 10. When one wishes to force a choice to gain a direction of attitude, a numerical scale may reduce the number of responses at

the extremes since it includes a middle position (see Mason [1991] for a summary of the middle position debate).

Devlin et al. (1993) express a preference for client satisfaction scales that embed expectations or ask whether the respondent requirements have been met. Only those who are completely satisfied in all respects (i.e., have had all their expectations met) will provide the highest rating. They report on split ballot experiments comparing various scales measuring client satisfaction.[4] Scales that measure the gap between expectations/requirements and performance have the highest reliability and validity and distribute the responses more widely compared to scales that only grade performance, which tend to cluster respondents in the highest category. Their conclusion is that unless expectations are included in context of the question, client satisfaction measurements are too optimistic.

Simply replacing a semantic scale with a numerical magnitude scale does not necessarily improve a question. The question meaning must be clear. The numerical scale is a good alternative to the conventional word-based response scale, with the following provisos. First, respondents must be comfortable with the concept of translating feelings and perceptions into a numerical scale. Second, the question context needs clarification. For example, in client satisfaction, expectations and requirements play an important role in mediating the experience of customers and must be factored into the question. Third, a 0–10 scale has an implicit benchmark of respondent experiences and services that have been a “0” or a “10” as well as those in between. It is useful to benchmark against previous experiences, provided that they are related. This is just another way of saying that calibration and context is needed with the numerical scale just as it is with physical scales such as pressure or light.

DECISION THEORY AND QUESTIONNAIRE DESIGN

The issue of context leads to a second theme in current questionnaire research: the influence that context has on question responses. Survey researchers have been very aware of the problems posed by inter-item contamination (questions influencing the responses following) and social desirability bias (respondents reluctant to reveal perceived unacceptable behaviour). In a client satisfaction survey, respondents may be reluctant to give a low rating if they believe that service providers are trying their best, or they may wish to compensate for a low rating or critical comment they made earlier.

Inter-item Contamination

Qa. In your view, is AIDS a threat to someone who is heterosexual and not a drug user?

Qb. Is the government providing sufficient funding to basic research in health?

Social Desirability Bias

Qc. Have you heard of the XYZ program?

Qd. In order to assess how well we are promoting our services, please tell me whether you have heard of the XYZ program?

Inter-item contamination arises from the serial nature of questions. Question Qa contaminates responses to question Qb, the extent of which can be determined in a split ballot where two versions of the questionnaire are tested, but the order of the questions is reversed.

Social desirability bias in this example stems from the natural reluctance we all have to admitting ignorance. Question Qc might encourage respondents to state knowledge they do not have. Question Qd provides a context where admitting ignorance is acceptable since it shifts the "blame" to failure with the program.

Both inter-item contamination and social desirability bias are manifestations of a deeper issue in questionnaire design. The traditional model of how a respondent provides an answer is termed the *accessibility hypothesis*. This approach views responses to questions as a process in which, after interpreting the question, a respondent searches a mental file system to seek the correct answer. In a simplistic view of this process, no difference is seen to exist among providing a response to a factual question about where one lives, a behavioural question on shopping patterns, or a values-based question about abortion. The respondent is viewed as having an instantly accessible file cabinet of answers to any question.

An alternate model called the *values-based approach* sees the respondent as engaged in a more strategic and judgemental process to evaluate appropriate responses before sharing them with the interviewer. From this perspective, questionnaire development confronts the respondent's value system; questions may be used, not so much to obtain information directly, but to create a context in which respondents are encouraged to reveal their true attitudes. According to this view, the underpinning of what we decide is good (or acceptable) is based on our concepts of "rationality."

The term “rationality” should not be confused with objective unemotional decisions or what people should do if they could perfectly forecast the future or could balance all alternatives objectively. In this context, rationality relates to consistency in applying certain rules. For example, if I said I prefer restaurant A to B, and also say I prefer B to C, then I should prefer restaurant A to C. Because subjectively valued alternatives are never precise, one can easily violate common rules of rationality. As issues and choices become more complex, such as on social issues or preferences about government intervention, it is common that subjective feelings are not rational in the strict sense. The term “bounded rationality” refers to decision making with multiple objectives, imperfect information, and constraints to action, all of which limit our ability to make the perfectly rational choice.

This leads to a key perception about the ability of respondents to be candid in the context of a brief relationship with a stranger (the interviewer). An everyday example is our usual response to the question “How are you?” Most of us choose not to burden our acquaintances and colleagues with a detailed review of our lives each time this question arises. We are strategic and judgemental in our response, knowing that even a slight intonation that qualifies the response “I am fine” could result in a prolonged and possibly unwanted interaction.

Survey respondents must often try to recall, judge, and comprehend within the context of the interview. They need to ensure that the responses made later in the interview are aligned (are rational) in the context of previous answers. It is common in political or philosophical debates with one’s peers or friends that logical flaws are revealed in our arguments. We try to patch these over. Respondents in an interview often try to do the same thing, and the easiest way to do this is to reveal only a limited view of one’s true feelings. Often, in-person interviews are favoured because they build a rapport designed to encourage respondents to reveal their true feelings. Others believe that a mailed questionnaire provides more assurance of confidential information being revealed simply because there is no interviewer present.

The idea that a rational approach minimizes information flow and conceals true feelings is not new. Most of the time researchers believed that respondents simply wish privacy. The recent focus on rational choice theory deals with the legitimate interests of

respondents to maintain consistency and the problems that a standardized interview has in allowing respondents to communicate a full perspective on their beliefs.

A simple illustration of this process is the tendency for respondents to keep their mental effort to a minimum. In a questionnaire context, respondents and researchers often minimize the effort needed to produce an acceptable (to the interviewer) outcome. Acceptability is usually determined by the willingness of the interviewer to continue to the next question without probing for clarification. Understanding that respondents are strategic and rational in their selection of responses illustrates the fallacy of most pretesting regimes that determine only whether the questionnaire "flows" smoothly. Respondents will naturally take the easy way out unless the researcher guides the process and insists that the respondent provide a considered response.

The idea that respondents provide answers that take less effort is illustrated when researchers argue that the middle position should *not* be offered. By forcing choice, respondents are required to exert more effort in selecting alternatives. The idea is that information becomes more precise when respondents are encouraged to consider and reconsider their responses.

There are deeper aspects to how respondents rationally frame their responses. Basic work has been done by Tversky and Kahnerman (1982), who developed the idea of a *decision frame* analogous to a visual perspective. This decision frame becomes the basis on which to understand choice; it sees respondents as rational actors who assess choices and preferences according to a subjective valuation of alternatives. The survey questionnaire should then set up response categories to mirror this choice procedure. Consider the following example.

Example (from Tversky and Kahnerman; 1982)
100 randomly selected respondents are each asked to make choices in the following situations:

Scenario 1
Imagine that Canada is preparing for the outbreak of an unusual disease that is expected to kill 600 people. Two alternative programs have been proposed with the following outcomes: [respondent preferences in brackets]

Program A 200 people will certainly be saved. [72 favoured this one]
Program B There is a 33% chance that 600 will be saved and a 67% chance that no one will be saved. [28 favoured this one]

Scenario 2
Same basic scenario as above, but with the following two program alternatives:

Program C 400 people will certainly die. [22 favoured this]
Program D 0 deaths with a 33% chance and 600 deaths with a 67% chance. [78 favoured this]

Strictly applying the rules of probability, Programs A and C have identical outcomes as do Programs B and D. For many respondents the equivalence of the two programs within each problem is lost in the different framing. This is encouraged by the negative framing of speaking of deaths instead of survivors and the fact that many interpret probability loosely. This overwhelms the ability of many respondents to distinguish among alternatives. The framing embedded within the questionnaire is sometimes subtle and influenced by everything said prior to it. Also critical is the notion that two events, one positive and the other negative, with probabilities of 33% and 67% respectively, are different than a single, certain, negative event. Underneath this example is the confusion that many have in dealing with probabilistic approaches to life-and-death situations. The frames determine our rational process and therefore influences our approach to selecting among what should be equal alternatives.

Framing in itself is neither good nor bad. It is a feature of questionnaires and the serial nature of communication that a response to a question is influenced by preceding questions and answers. The use of framing can be very important to creating a base of information from which the respondent is equipped to respond to complex issues. Again, the use of split ballots can increase one's understanding of a complex issue. Of course, framing can also be used to push or pull the survey respondents in one or the other direction. This latter tendency is becoming more evident in political polling.

Tversky and Kahnerman (1982) see choice as the outcome of two phases: initial framing of acts and contingencies, and evaluation and choice. Using the type of problem above, the responses can be shifted

dramatically by changing the frame of the problem. Clearly, it makes little sense for Program C to be less favoured than Program A: they are equivalent.

When we administer a standardized questionnaire, we ask respondents to perform five basic information tasks:

- provide facts (age, income, number in household, etc.)
- divulge belief and knowledge (Is Elvis alive?)
- judge outcomes or states (If X were leader of the PCs, which Liberal politician would you choose for leader of that party?)
- provide responses to hypothetical situations (Whom would you vote for as Prime Minister - X or Y?)
- evaluate degree of goal achievement (Overall, how well has the government done with the economy?)

Understanding the cognitive basis for question comprehension is basic to improved questionnaire design. Four cognitive steps are involved in respondents providing valid and reliable information:

- respondent understanding of question intent (*comprehension*);
- respondent ability/willingness to access the required information (*accessibility*);
- respondent capacity to formulate a response based on the information retrieved (*retrieval*); and
- respondent translation of the choice he or she has made into the response categories provided (*communication*).

Cognitive science is beginning to explore how these four factors influence questionnaire data. Some recent examples of this work are found in Alwin (1991) and Esser (1990). It is important to note that the traditional accessibility model of response delivery on a standardized questionnaire is limited. Respondents perform complex mental tasks to answer even simple questions. From this perspective, the context of questions, providing clues and cues and other mental prompts are essential features of the design of a questionnaire. This view is juxtaposed with common practice, in which questionnaires need to amass facts within time constraints and every question must contribute to substantive issues. In an evaluation context, researchers have tried to show clients how each question on the survey is linked directly to substantive issues underlying the evaluation. Questions designed to “set the stage” are often discarded as too expen-

sive or a waste. The central problem facing the evaluator is to prove to managers that these context questions materially enhance other substantive questions. This area is an exciting new field of research into questionnaire design.

Three practical suggestions emerge from the literature. First, evaluators need to use split ballots to test their questions and question sequences for order effects and inter-item contamination. Second, where order effects are demonstrated, preliminary "stage setting" questions are needed to ensure that respondents understand the concepts. Third, additional resources are needed at the design phase. More than just ensuring that the questionnaire "flows," researchers need to ensure that respondents understand the concepts and are not being induced to respond in a certain way. Focus groups are a useful method of pre-testing a questionnaire. Another approach is to debrief pre-test respondents on each question to probe their understanding, and to discover areas in which they might be reluctant to respond. In an era of constrained research budgets and managers intent on confining resources only to collecting "hard" data, these suggestions are hard to fulfil. Yet, without careful design, the reliability and validity of evaluation work will be substantially compromised.

QUESTIONNAIRE FRAMEWORKS TO REPLICATE RESPONDENT DECISION MAKING

The third area in which questionnaire design has evolved reflects a shift in the metaphor of the standardized questionnaire. Since the early 1960s, market research literature has expressed growing discomfort over the way questionnaires were used to collect information on intention to purchase and sensitivity to price. Consider the example of determining what a car purchaser wishes. One can enumerate the options and prices, building an increasingly loaded vehicle. Once the total price is announced, the consumer often decides not to purchase. Is there a method for testing willingness-to-pay and likelihood of purchasing while observing the total budget? Can we design a questionnaire that replicates the consumer choice process?

Increasingly, questionnaires are frameworks or models of consumer decision making and place the consumer in circumstances resembling the product choice process. In this view, the question text and their sequence is determined by an experimental design and a view of how consumers decide to purchase. Here the context and sequence

of questions is used to increase the validity and reliability of the information collected.

Conjoint analysis or *contingent valuation* are two methods commonly used to develop a more consistent picture of preferences for new products and services, policy evaluation, and assessing public programs. Determining the "value" of a program to users is a typical example of where these ideas can be applied by the evaluator. These methodologies constitute a specific approach to questionnaire design along with an associated statistical analysis procedure. Only the questionnaire design aspects are relevant here.

Conjoint analysis views products or services (public and private) as comprising a bundle of *attributes* (see Louviere [1988] or Green [1984]). Users (clients) choose among services by simultaneously weighing attributes (including cost) to maximize their net welfare. The standardized questionnaire encourages respondents to see product/service attributes as sequential, which is why results can be so inconsistent (as in the car example). Conjoint analysis reveals what combinations of attributes are preferred by the client when all attributes and costs are considered simultaneously.

Consider education and training programs. Such programs could be described using three attributes: academics, job search skills, and life skills training. Within each attribute conjoint analysis requires that at least two levels be defined. For academics one might identify three levels such as grade 8 (level 1), grade 10 (level 2), and high school equivalency (level 3). Both levels and attributes must be easily comprehended by survey respondents for this approach to be successful. These attributes would then be tested using a questionnaire that provides a respondent with a number of alternative packages to rate. A typical package, reflecting a level of each attribute might be:

> *Package 1*
> On the following scale of 0–10, please rank the following educational package:
> Academics - Grade 8 (Level 1)
> Job Search Skills - Resume writing (Level 2)
> Life Skills - None (Level 1)

Notice there is a low level for attribute 1, a mid level for attribute 2, and a low level for attribute 3. Another question could vary the levels of each attribute.

Package 2
On the following scale of 0–10, please rank the following educational package:
Academics - Grade 10 (Level 2)
Job Search Skills - Resume writing (Level 2)
Life Skills - Assertiveness, organization, motivation courses (Level 3)

If each attribute has three levels and there are three attributes, 27 different (3^3) combinations of packages exist. An obvious constraint is that respondents usually have limits to the number of packages they can rate. Using experimental design methods, one can distribute the different attributes and levels across fewer packages and maintain a viable data set on which to assess the value of each attribute. One tries to examine how the rating scale varies with changes in the levels of each attribute. A regression model is a common analysis approach in which the scale value is regressed on dummy independent variables that indicate the presence or absence of a given attribute level. Green (1984) provides a good survey of this method.

The description of a specific level within an attribute must be exactly the same in different packages, such as "resume writing" in the two packages above. A conjoint questionnaire starts with questions that ask respondents to rate the attributes and levels. Each level is rated on a 1–5 scale and respondents are asked to distribute 100 votes or dollars among the attributes. This information is used in some statistical procedures to calibrate the estimation process, but it also serves to set the stage for the rating of the packages that forms the core of the questionnaire. In this sense conjoint techniques observe the guidelines for creating a context or decision frame outlined above. Typical conjoint questionnaires ask respondents to rate 10 packages. The final section of a conjoint questionnaire collects personal information to act as covariates on the regression models. The basic point is that the format of the questionnaire and the form of the questions are dictated by an experimental design that is very similar to those used to allocate subjects in medical and psychological experiments.

A key requirement for this technique to work is that attributes and levels must be evocative and clear. For example, to describe the three levels under academics as "elementary," "intermediate," and "advanced" is too vague. Descriptions such as "grade 8," "grade 10," and "high school equivalency" allow the respondent to understand

what is being offered in a package. In other words, the attributes and levels must be grounded.

Conjoint analysis is used in market research to determine the composition of attributes that need to be included in a product or service, while simultaneously including a price constraint in the decision. To use the car example, the consumer would value packages of attributes (engine size, style, etc.) at a given price. The rating of packages reflects the levels within each attribute that are valued, and this guides the manufacturer in the creation of the product. This technique has an important place in the commercialization of public services. For example, many services, such as information products from Statistics Canada or the one-stop shopping initiatives of the federal government, have been offered without regard to the value perceived by the consumer. Conjoint analysis can provide very useful insight into the appropriate pricing and product mix that should be offered.

Contingent valuation is used to price non-market goods and services typical of government. This approach measures the value of a program by testing what people are willing to pay to retain the service. An equivalent approach is to measure how much compensation people need to tolerate its removal. The following example is drawn from research completed for Manitoba Hydro by the author (Mason, 1995).

Example: The Value of Preserving Wilderness

Manitoba Hydro needed to carry power from its new northern development south to Ontario. Two alternative transmission routes existed. One option was to run the new transmission line along an established route close to existing hydro lines. An alternate and more direct (cheaper) route was through wilderness areas. The object of this research was to place a money value on wilderness.

Two questions were posed to ratepayers:

- What are you willing to pay extra on your monthly bill to use the existing, more costly route and avoid having the transmission line go through the wilderness area?

- What reduction in your hydro bill would you accept to have the transmission line go through the wilderness area?

Respondents were enrolled in a telephone interview and sent a letter informing them of the issues. A map outlined the features of each alternative route. The cost differential in the two routes was not mentioned, because the focus was to place a monetary value on wilderness. To simplify a complex design, some respondents were asked in a follow-up telephone interview whether they would be willing to pay "X" as a surcharge on their monthly utility bill to use Option 1 (more costly established route) compared to constructing a cheaper line, but through wilderness. Others were asked whether they would accept a reduction of "X" if the new line were constructed through wilderness ("X" was a fraction calibrated to the respondent's monthly electric bill).

The result of this research showed that hydro ratepayers were willing to pay enough to run the new line along the existing route. In others words, they were willing to pay a surcharge that totalled about $25 million over 25 years to preserve wilderness. This example shows the power of contingent valuation in assessing the value of various policy initiatives. Clearly, this approach can be used in a range of public programs.

Regulation and policies related to gambling, public health, and so on can all be assessed using this willingness-to-pay model. As fiscal restraint becomes more binding, evaluators will be asked to help place a price on previously free public programs.

CONCLUSION

The basic rules for questionnaire design still provide reasonable guidance for many evaluations. Recent developments in scaling, advances in concepts of rationality, and new frameworks for modelling respondent choice behaviour are deepening our understanding of the complexities underlying any standardized questionnaires. This paper has attempted to illustrate some recent developments in questionnaire design. The one central theme from this new research is that questionnaires are much more complex than many evaluators believe.

NOTES

1. A standardized questionnaire attempts to pose exactly the same questions to a relatively large sample of respondents. There is little room for accepting interpretation by the respondent.

2. A monotonic scale allows only one category per respondent. In other words, a respondent cannot check a "3" and "8" for the same question. It is linear in that the difference between a 5 and a 6 is the same as the difference between an 8 and a 9.

3. Technically, this use of a numerical scale is not correct, because a numerical scale in the 0–10 form is not a ratio scale, but an interval scale. Most researchers gloss over this.

4. A split ballot refers to a process of administering different form of the questionnaire to randomly selected subsets of the sample. This allows a researcher to determine whether the different question wordings produce variation in the patterns of response.

REFERENCES

Alwin, D.F. (1991). Information transmission in the survey interview: Number of response categories and reliability of attitude measurement. *Sociological Methodology, 21*(3), 3–27.

Belson, A. (1981). *The design and understanding of survey questions*. Aldershot, UK: Gower.

Converse, J.M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire* (Sage University Paper 63). Newbury Park, CA: Sage.

Devlin, S., Dong, J.H.K, & Brown, M. (1993). Selecting a scale for measuring quality. *Marketing Research, 5*(3), 5–16.

Esser, H. (1990). *Response set: Habit, frame or rational choice*. Paper presented at Relevance of Attitude Measurement in Sociology conference, Bad Homburg, Germany (cited in Alwin, 1991).

Green, P.E. (1984). Hybrid models for conjoint analysis: An expository review. *Journal of Marketing Research, 22*, 155–189.

Lodge, M. (1981). *Magnitude scaling* (University Paper Number 25). Berkeley, CA: Sage.

Louviere, J.J. (1988). *Analyzing decision making: Metric conjoint analysis*. Newbury Park, CA: Sage.

Mason, G.C. (1991). Issues in designing the standardized questionnaire. In A. Love (Ed.), *Evaluation methods sourcebook* (pp. 26–43). Ottawa: Canadian Evaluation Society.

Mason, G.C. (1995). *The value of wilderness* (Working Paper). Winnipeg: University of Manitoba, Department of Economics.

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

Sheatsley, P.R. (1983). Questionnaire construction and item writing. In P.H. Rossi, J.D. Wright, & A.B. Anderson (Eds.), *Handbook of survey research* (pp. 195–230). New York: Academic Press.

Tversky, A., & Kahnerman, D. (1982). The framing of decisions and the psychology of choice. In R. Hogarth (Ed.), *Question framing and response consistency: New directions for methodology of social and behavioral science, 11* (pp. 56–78). San Francisco: Jossey-Bass.

Waddell, H. (1995). Getting a straight answer. *Marketing Research, 7*(3), 4–8.