A Brief Guide to Questionnaire Development

Robert B. Frary
Office of Measurement and Research Service
Virginia Polytechnic Institute and State University

Introduction

Most people have responded to so many questionnaires in their lives that they have little concern when it becomes necessary to construct one of their own. Unfortunately the results are often unsatisfactory. One reason for this outcome may be that many of the questionnaires in current use have deficiencies which are consciously or unconsciously incorporated into new questionnaires by inexperienced developers. Another likely cause is inadequate consideration of aspects of the questionnaire process separate from the instrument itself, such as how the responses will be analyzed to answer the related research questions or how to account for nonreturns from a mailed questionnaire.

These problems are sufficiently prevalent that numerous books and journal articles have been written addressing them (e.g., see Dillman, 1978). Also, various educational and proprietary organizations regularly offer workshops in questionnaire development. Therefore, this booklet is intended to identify some of the more prevalent problems in questionnaire development and to suggest ways of avoiding them. This paper does not cover the development of inventories designed to measure psychological constructs, which would require a deeper discussion of psychometric theory than is feasible here. Instead, the focus will be on questionnaires designed to collect factual information and opinions.

Preliminary Considerations

Some questionnaires give the impression that their authors tried to think of every conceivable question that might be asked with respect to the general topic of concern. Alternatively, a committee may have incorporated all of the questions generated by its members. Stringent efforts should be made to avoid such shotgun approaches, because they tend to yield very long questionnaires often with many questions relevant to only small proportions of the sample. The result is annoyance and frustration on the part of many responders. They resent the time it takes to answer and are likely to feel their responses are unimportant if many of the questions are inapplicable. Their annoyance and frustration then causes nonreturn of mailed questionnaires and incomplete or inaccurate responses on questionnaires administered directly. These difficulties can yield largely useless results. Avoiding them is relatively simple but does require some time and effort.

The first step is mainly one of mental discipline. The investigator must define precisely the information desired and endeavor to write as few questions as possible to obtain it. Peripheral questions and ones to find out "something that might just be nice to know" must be avoided. The author should consult colleagues and potential consumers of the results in this process.

A second step, needed for development of all but the simplest questionnaires, is to obtain feedback from a small but representative sample of potential responders. This activity may involve no more than informal, open-ended interviews with several potential responders. However, it is better to ask such a group to criticize a preliminary version of the questionnaire. In this case, they should first answer the questions just as if they were research subjects. The purpose of these activities is to determine relevance of the questions and the extent to which there may be problems in obtaining responses. For example, it might be determined that responders are likely to be offended by a certain type of question or that a line of questions misconstrues the nature of a problem the responders encounter.

The process just described should not be confused with a field trial of a tentative version of the questionnaire. This activity also is desirable in many cases but has different purposes and should always follow the more informal review process just described. A field trial will be desirable or necessary if there is substantial uncertainty in areas such as:

1) Response rate. If a field trial of a mailed questionnaire yields an unsatisfactory response rate, design changes or different data gathering procedures must be undertaken.

2) Question applicability. Even though approved by reviewers, some questions may prove redundant. For example, everyone or nearly everyone may be in the same answer category for some questions, thus making them unnecessary.

3) Question performance. The field-trial response distributions for some questions may clearly indicate that they are defective. Also, pairs or sequences of questions may yield inconsistent responses from a number of trial responders, thus indicating the need for rewording or changing the response mode.

Writing the Questionnaire Items

Open-Ended Questions

While these seem easy to write, in most cases they should be avoided. A major reason is variation in willingness and ability to respond in writing. Unless the sample is very homogeneous with respect to these two characteristics, response bias is likely. Open-ended questions are quite likely to suppress responses from the less literate segments of a population or from responders who are less concerned about the topic at hand.

A reason frequently given for using open-ended questions is the capture of unsuspected information. This reason is valid for brief, informal questionnaires to small groups, say, ones with fewer than 50 responders. In this case, a simple listing of the responses to each question usually conveys their overall character. However, in the case of a larger sample, it is necessary to categorize the responses to each question in order to analyze them. This process is time-consuming and introduces error. It is far better to determine the prevalent categories in advance and ask the responders to select among those offered. In most cases, obscure categories applicable only to very small minorities of responders should not be included. A preliminary, open-ended questionnaire sent to a small sample is often a good way to establish the prevalent categories in advance.

Contrary to the preceding discussion, there are circumstances under which it may be better to ask the responders to fill in blanks. This is the case when the responses are to be hand entered into computer data sets and when the response possibilities are very clearly limited and specific. For example, questions concerning age, state of residence, or credit-hours earned may be more easily answered by filling in blanks than by selecting among categories. If the answers are numerical, this response mode may also enhance the power of inferential statistical procedures. If handwritten answers are to be assigned to categories for analysis, flexibility in category determination becomes possible. However, if the responders are likely to be estimating their answers, it is usually better to offer response categories (e.g., to inquire about body weight, grade-point average, annual income, or distance to work).

Objective Questions

With a few exceptions, the category "Other" should be avoided as a response option, especially when it occurs at the end of a long list of fairly lengthy choices. Careless responders will overlook the option they should have designated and conveniently mark the option "other." Other responders will be hairsplitters and will reject an option for some trivial reason when it really applies, also marking "other." "Other (specify)" or "other (explain)" may permit recoding these erroneous responses to the extent that the responders take the trouble to write coherent explanations, but this practice is time-consuming and probably yields no better results than the simple omission of "other." Of course, the decision not to offer the option "other" should be made only after a careful determination of the categories needed to classify nearly all of the potential responses. Then, if a few responders find that, for an item or two, there is no applicable response, little harm is done.

An exception to the foregoing advice is any case in which the categories are clear-cut, few in number, and such that some responders might feel uncomfortable in the absence of an applicable response. For example, if nearly all responders would unhesitatingly classify themselves as either black or white, the following item would serve well:

```
Race:   1) Black   2) White   3) Other
```

Also consider:

```
Source of automobile: 1) Purchased new  2) Purchased used
                      3) Other
```

"Other (specify)" should be used only when the investigator has been unable to establish the prevalent categories of response with reasonable certainty. In this case, the investigator is clearly obligated to categorize and report the "other" responses as if the question were open-ended. Often the need for "other" reflects inadequate efforts to determine the categories that should be offered.

Issues

Category Proliferation.

A typical question is the following:

```
Marital status:  1) Single (never married)     4) Divorced
                 2) Married                    5) Separated
                 3) Widowed
```

Unless the research in question were deeply concerned with conjugal relationships, it is inconceivable that the distinctions among all of these categories could be useful. Moreover, for many samples, the number of responders in the latter categories would be too small to permit generalization. Usually, such a question reflects the need to distinguish between a conventional familial setting and anything else. If so, the question could be:

```
Marital status:  1) Married and living with spouse
                 2) Other
```

In addition to brevity, this has the advantage of not appearing to pry so strongly into personal matters.

Scale Point Proliferation.

In contrast to category proliferation, which seems usually to arise somewhat naturally, scale point proliferation takes some thought and effort. An example is:

```
1) Never  2) Rarely  3) Occasionally  4) Fairly often
5) Often  6) Very often 7) Almost always  8) Always
```

Such stimuli run the risk of annoying or confusing the responder with hairsplitting differences between the response levels. In any case, psychometric research has shown that most subjects cannot reliably distinguish more than six or

seven levels of response, and that for most scales a very large proportion of total score variance is due to direction of choice rather than intensity of choice. Offering four to five scale points is usually quite sufficient to stimulate a reasonably reliable indication of response direction.

Questionnaire items that ask the responder to indicate strength of reaction on scales labeled only at the end points are not so likely to cause responder antipathy if the scale has six or seven points. However, even for semantic differential items, four or five scale points should be sufficient.

Order of Categories.

When response categories represent a progression between a lower level of response and a higher one, it is usually better to list them from the lower level to the higher in left-to-right order, for example,

1) Never 2) Seldom 3) Occasionally 4) Frequently

This advice is based only on anecdotal evidence, but it seems plausible that associating greater response levels with lower numerals might be confusing for some responders.

Combining Categories.

In contrast to the options listed just above, consider the following:

1) Seldom or never 2) Occasionally 3) Frequently

Combining "seldom" with "never" might be desirable if responders would be very unlikely to mark "never" and if "seldom" would connote an almost equivalent level of activity, for example, in response to the question, "How often do you tell you wife that you love her?" In contrast, suppose the question were, "How often do you drink alcoholic beverages?" Then the investigator might indeed wish to distinguish those who never drink. When a variety of questions use the same response scale, it is usually undesirable to combine categories.

Responses at the Scale Midpoint.

Consider the following questionnaire item:

The instructor's verbal facility is:

```
1) Much below average          4) Above average
2) Below average               5) Much above average
3) Average
```

Associating scale values of 1 through 5 to these categories can yield highly misleading results. The mean for all instructors on this item might be 4.1, which, possibly ludicrously, would suggest that the average instructor was above average. Unless there were evidence that most of the instructors in question were actually better than average with respect to some reference group, the charge of using statistics to create false impressions could easily be raised.

A related difficulty arises with items like:

The instructor grades fairly.

```
1) Agree              4) Tend to disagree
2) Tend to agree      5) Disagree
3) Undecided
```

There is no assurance whatsoever that a subject choosing the middle scale position harbors a neutral opinion. A subject's choice of the scale midpoint may result from:

Ignorance--the subject has no basis for judgment.

Uncooperativeness--the subject does not want to go to the trouble of formulating an opinion.

Reading difficulty--the subject may choose "Undecided" to cover up inability to read.

Reluctance to answer--the subject may wish to avoid displaying his/her true opinion.

Inapplicability--the question does not apply to the subject.

In all the above cases, the investigator's best hope is that the subject will not respond at all. Unfortunately, the seemingly innocuous middle position counts, and, when a number of subjects choose it for invalid reasons, the average response level is raised or lowered erroneously (unless, of course, the mean of the valid responses is exactly at the scale midpoint).

The reader may well wonder why neutral response positions are so prevalent on questionnaires. One reason is that, in the past, crude computational methods were unable to cope with missing data. In such cases, nonresponses were actually replaced with neutral response values to avoid this problem. The need for such a makeshift solution has long been supplanted by improved computational methods, but the practice of offering a neutral response position seems to have a life of its own. Actually, if a substantial proportion of the responders really do hold genuinely neutral opinions and will cooperate in revealing these, scale characteristics will be enhanced modestly by offering a neutral position. However, in most cases, the potential gain is not worth the risk.

In the absence of a neutral position, responders sometimes tend to resist making a choice in one direction or the other. Under this circumstance, the following strategies may alleviate the problem:

1) Encourage omission of a response when a decision cannot be reached.

2) Word responses so that a firm stand may be avoided, e.g., "tend to disagree."

3) If possible, help responders with reading or interpretation problems, but take care to do so impartially and carefully document the procedure so that it may be inspected for possible introduction of bias.

4) Include options explaining inability to respond, such as "not applicable," "no basis for judgment," "prefer not to answer."

The preceding discussion notwithstanding, there are some items that virtually require a neutral position. Examples are:

How much time do you spend on this job now?

```
1) Less than before  2) About the same  3) More time
```

The amount of homework for this course was

```
1) too little.  2) reasonable.  3) too great.
```

It would be unrealistic to expect a responder to judge a generally comparable or satisfactory situation as being on one side or another of the scale midpoint.

Response Category Language and Logic

The extent to which responders agree with a statement can be assessed adequately in many cases by the options:

```
1) Agree      2) Disagree
```

However, when many responders have opinions that are not very strong or well-formed, the following options may serve better:

```
1) Agree   2) Tend to agree   3) Tend to disagree
4) Disagree
```

These options have the advantage of allowing the expression of some uncertainty.

In contrast, the following options would be undesirable in most cases:

```
1) Strongly agree  2) Agree  3) Disagree
4) Strongly Disagree
```

While these options do not bother some people at all, others find them objectionable. "Agree" is a very strong word; some would say that "Strongly agree" is redundant or at best a colloquialism. In addition, there is no comfortable resting place for those with some uncertainty. There is no need to unsettle a segment of responders by this or other cavalier usage of language.

Another problem can arise when a number of questions all use the same response categories. The following item is from an actual questionnaire:

Indicate the extent to which each of the following factors influences your decision on the admission of an applicant: Amount of Influence

	None	Weak	Moder	Strong
SAT/ACT scores	_____	_____	_____	_____
High school academic record	_____	_____	_____	_____
Extracurricular activities	_____	_____	_____	_____
Personal interview	_____	_____	_____	_____
Open admissions	_____	_____	_____	_____

Only sheer carelessness could have caused failure to route the responder from a school with open admissions around the questions concerning the influence of test scores, etc. This point aside, consider the absurdity of actually asking a responder from an open admissions school to rate the influence of their open admissions policy. (How could it be other than strong?) Inappropriate response categories and nonparallel stimuli can go a long way toward inducing disposal rather than return of a questionnaire.

A subtle but prevalent error is the tacit assumption of a socially conventional interpretation on the part of the responder. Two examples from actual questionnaires are:

Indicate how you feel about putting your loved one in a nursing home.

1) Not emotional 2) Somewhat emotional
3) Very emotional

How strong is the effect of living at some distance from your family?

1) Weak 2) Moderately strong 3) Very strong

Obviously (from other content of the two questionnaires), the investigators never considered that many people enjoy positive emotions upon placing very sick individuals in nursing homes or beneficial effects due to getting away from troublesome families. Thus, marking the third option for either of these items could reflect either relief or distress, though the investigators interpreted these responses as indicating only distress. Options representing a range of positive to negative feelings would resolve the problem.

A questionnaire from a legislative office used the following scale to rate publications:

1) Publication legislatively mandated
2) Publication not mandated but critical to agency's effectiveness
3) Publication provides substantial contribution to agency's effectiveness
4) Publication provides minor contribution to agency's effectiveness

This is a typical example of asking two different questions with a single item, namely: a) Was the publication legislatively mandated? and b) What contribution did it make? Of course, the bureaucrats involved were assuming that any legislatively mandated publication was critical to the agency's effectiveness. Note that options 3 and 4 but not 2 could apply to a mandated publication, thus raising the possibility of (obviously undesired) multiple responses with respect to each publication.

Ranking Questions

Asking responders to rank stimuli has drawbacks and should be avoided if possible. Responders cannot be reasonably expected to rank more than about six things at a time, and many of them misinterpret directions or make mistakes in responding. To help alleviate this latter problem, ranking questions may be framed as follows:

Following are three colors for office walls:

```
     1) Beige  2) Ivory  3) Light green

Which color do you like best?          _____
Which color do you like second best?   _____
Which color do you like least?         _____
```

The "Apple Pie" Problem

There is sometimes a difficulty when responders are asked to rate items for which the general level of approval is high. For example, consider the following scale for rating the importance of selected curriculum elements:

```
1) No importance    3) Moderate importance
2) Low importance   4) High importance
```

Responders may tend to rate almost every curriculum topic as highly important, especially if doing so implies professional approbation. Then it is difficult to separate topics of greatest importance from those of less. Asking responders to rank items according to importance in addition to rating them will help to resolve this problem. If there are too many items for ranking to be feasible, responders may be asked to return to the items they have rated and indicate a specified small number of them that they consider "most important."

Another strategy for reducing the tendency to mark every item at the same end of the scale is to ask responders to rate both positive and negative stimuli. For example:

```
My immediate supervisor:

  handles employee problems well.       1) Agree  2) Disagree
  works with us to get the job done.    1) Agree  2) Disagree
  embarrasses those who make mistakes.  1) Agree  2) Disagree
  is a good listener                    1) Agree  2) Disagree
  often gives unclear instructions      1) Agree  2) Disagree
```

Flatfooted negation of stimuli that would normally be expressed positively should be avoided when this strategy is adopted. For example, "does not work with us to get the job done" would not be a satisfactory substitute for the second item above.

Unnecessary Questions

A question like the following often appears on questionnaires sent to samples of college students:

```
Age:  1) below 18   2) 18-19   3) 20-21   4) over 21
```

If there is a specific need to generalize results to older or younger students, the question is valid. Also, such a question might be included to check on the representativeness of the sample. However, questions like this are often included in an apparently compulsive effort to characterize the sample exhaustively. A clear-cut need for every question should be established. This is especially important with respect to questions characterizing the responders, because there may be a tendency to add these almost without thought after establishment of the more fundamental questions. The fact that such additions may lengthen the questionnaire needlessly and appear to pry almost frivolously into personal matters is often overlooked. Some questionnaires ask for more personal data than opinions on their basic topics.

In many cases, personal data are available from sources other than the responders themselves. For example, computer files used to produce mailing labels often have other information about the subjects that can be merged with their responses if these are not anonymous. In such cases, asking the responders to repeat this information is not only burdensome but may introduce error, especially when reporting the truth has a negative connotation. (Students often report inflated grade-point averages on questionnaires.)

Sensitive Questions

When some of the questions that must be asked request personal or confidential information, it is better to locate them at the end of the questionnaire. If such questions appear early in the questionnaire, potential responders may become too disaffected to continue, with nonreturn the likely result. However, if they reach the last page and find unsettling questions, they may continue nevertheless or perhaps return the questionnaire with the sensitive questions unanswered. Even this latter result is better than suffering a nonreturn.

Statistical Considerations

It is not within the scope of this booklet to offer a discourse on the many statistical procedures that can be applied to analyze questionnaire responses. However, it is important to note that this step in the overall process cannot be divorced from the other development steps. A questionnaire may be well-received by critics and responders yet be quite resistant to analysis. The method of analysis should be established before the questions are written and should direct their format and character. If the developer does not know precisely how the responses will be analyzed to answer each research question, the results are in jeopardy. This caveat does not preclude exploratory data analysis or the emergence of serendipitous results, but these are procedures and outcomes that cannot be depended on.

In contrast to the lack of specificity in the preceding paragraph, it is possible to offer one principle of questionnaire construction that is generally helpful with respect to subsequent analysis. This is to arrange for a manageable number of ordinally scaled variables. A question with responses such as:

```
1) Poor  2) Fair  3) Good  4) Excellent
```

will constitute one such variable, since there is a response progression from worse to better (at least for almost all speakers of English).

In contrast, to the foregoing example, consider the following question:

Which one of the following colors do you prefer for your office wall?

```
1) Beige  2) Ivory  3) Light green
```

There is no widely-agreed-upon progression from more to less, brighter to duller, or anything else in this case. Hence, from the standpoint of scalability, this question must be analyzed as if it were three questions (though, of course, the responder sees only the single question):

```
Do you prefer beige?              1) yes  2) no
Do you prefer ivory?              1) yes  2) no
Do you prefer light green?        1) yes  2) no
```

These variables (called dummy variables) are ordinally scalable and are appropriate for many statistical analyses. However, this approach results in proliferation of variables, which may be undesirable in many situations, especially those in which the sample is relatively small. Therefore, it is often desirable to avoid questions whose answers must be scaled as multiple dummy variables. Questions with the instruction "check all that apply" are usually of this type. (See also the comment about "check all that apply" under Optical Mark Reader Processing of Responses below).

Anonymity

For many if not most questionnaires, it is necessary or desirable to identify responders. The commonest reasons are to check on nonreturns and to permit associating responses with other data on the subjects. If such is the case, it is a clear violation of ethics to code response sheets surreptitiously or secretly to identify responders after stating or implying that responses are anonymous. In so doing, the investigator has in effect promised the responders that their responses cannot be identified. The very fact that at some point the responses can be identified fails to provide the promised security, even though the investigator intends to keep them confidential.

If a questionnaire contains sensitive questions yet must be identified for accomplishment of its purpose, the best policy is to promise confidentiality but not anonymity. In this case a code number should be clearly visible on each copy of the instrument, and the responders should be informed that all responses will be held in strict confidence and used only in the generation of statistics. Informing the responders of the uses planned for the resulting statistics is also likely to be helpful.

Nonreturns

The possibilities for biasing of mailed questionnaire results due to only partial returns are all too obvious. Nonreturners may well have their own peculiar views toward questionnaire content in contrast to their more cooperative co-recipients. Thus it is strange that very few published accounts of questionnaire-based research report any attempt to deal with the problem. Some do not even acknowledge it.

There are ways of at least partially accounting for the effects of nonreturns after the usual follow-up procedures, such as postcard reminders. To the extent that responders are asked to report personal characteristics, those of returners may be compared to known population parameters. For example, the proportion of younger returners might be much smaller than the population proportion for people in this age group. Then results should be applied only cautiously with respect to younger individuals. Anonymous responses may be categorized according to postal origin (if mailed). Then results should be applied more cautiously with respect to under represented areas.

Usually, the best way to account for nonresponders is to select a random sample of them and obtain responses even at substantial cost. This is possible even with anonymous questionnaires, though, in this case, it is necessary to contact recipients at random and first inquire as to whether they returned the questionnaire. Telephone interviews are often satisfactory for obtaining the desired information from nonresponders, but it is almost always necessary to track down some nonresponders in person. In either case, it may not be necessary to obtain responses to all questionnaire items. Prior analyses may reveal that only a few specific questions provide a key to a responder's opinion(s).

Format and Appearance

It seems obvious that an attractive, clearly printed and well laid out questionnaire will engender better response than one that is not. Nevertheless, it would appear that many investigators are not convinced that the difference is worth the trouble. Research on this point is sparse, but experienced investigators tend to place considerable stress on extrinsic characteristics of questionnaires. At the least, those responsible for questionnaire development should take into consideration the fact that they are representing themselves and their parent organizations by the quality of what they produce.

Mailed questionnaires, especially, seem likely to suffer nonreturn if they appear difficult or lengthy. A slight reduction in type size and printing on both sides of good quality paper may reduce a carelessly arranged five pages to a single sheet of paper.

Obviously, a stamped or postpaid return envelope is highly desirable for mailed questionnaires. Regardless of whether an envelope is provided, a return address should be prominently featured on the questionnaire itself.

Optical Mark Reader Processing of Responses

If possible, it is highly desirable to collect questionnaire responses on sheets that can be machine read. This practice saves vast amounts of time otherwise spent keying responses into computer data sets. Also, the error rate for keying data probably far outstrips the error rate of responders due to misplaced or otherwise improper marks on the response sheets.

Obtaining responses directly in this manner is almost always feasible for group administrations but may be problematical for mailed questionnaires, especially if the questions are not printed on the response sheet. Relatively unmotivated responders are unlikely to take the trouble to obtain the correct type of pencil and figure out how to correlate an answer sheet with a separate set of questions. Some investigators enclose pencils to motivate responders.

On the other hand, machine readable response sheets with blank areas, onto which questions may be printed, are available. Also, if resources permit, custom machine-readable sheets can be designed to incorporate the questions and appropriate response areas. The writer knows of no evidence that return rates suffer when machine readable sheets with the questions printed on them are mailed. Anecdotally, it has been reported that responders may actually be more motivated to return machine readable response sheets than conventional instruments. This may be because they believe that their responses are more likely to be counted than if the responses must be keyed. (Many investigators know of instances where only a portion of returned responses were keyed due to lack of resources.) Alternatively, responders may be mildly impressed by the technology employed or feel a greater degree of anonymity. In planning for the use of a mark reader, it is very important to coordinate question format with reader capability and characteristics. This coordination should also take planned statistical analyses into consideration. Questions that need to be resolved in the development phase include:

1. What symbolic representation (in a computer readable data set) will the various response options have (e.g., numerals, letters, etc.)?
2. How will nonresponse to an item be represented?
3. How will non-codeable responses (e.g., double marks) be represented?

Most readers are designed (or programmed) to recognize only a single intended answer to a given question. Given the ubiquity of "mark all that apply" instructions in questionnaires, it is therefore necessary to modify such questions for machine-compatible responding. The following example shows how this may be accomplished:

12. In which of these leisure activities do you participate at least once a week (check all that apply):

Swimming _____
Gardening _____
Golf _____
Bicycling _____
Tennis _____
Jogging _____

Questions 12-17 are a list of leisure activities. Indicate whether you participate in each activity at least once a week.

12. Swimming 1) Yes 2) No
13. Gardening 1) Yes 2) No
14. Golf 1) Yes 2) No
15. Bicycling 1) Yes 2) No
16. Tennis 1) Yes 2) No
17. Jogging 1) Yes 2) No

This procedure creates dummy variables suitable for many statistical procedures (see Statistical Considerations above).

Folding response sheets for mailing may cause processing difficulties. Folding may cause jams in the feed mechanisms of some readers. Another problem is that the folds may cause inaccurate reading of the responses. In these cases, sheet-size envelopes may be used for sending and return. Some types of opscan sheets can be folded, however, and these may be sent in business-size envelopes.

Sample Size

Various approaches are available for determining the sample size needed for obtaining a specified degree of accuracy in estimation of population parameters from sample statistics. All of these methods assume 100% returns from a random sample. (See Hinkle, Oliver, and Hinkle, 1985.)

Random samples are easy to mail out but are virtually never returned at the desired rate. It is possible to get 100% returns from captive audiences, but in most cases these could hardly be considered random samples. Accordingly, the typical investigator using a written questionnaire can offer only limited assurance that the results are generalizable to the population of interest. One approach is to obtain as many returns as the sample size formulation calls for and offer evidence to show the extent of adherence of the obtained sample to known population characteristics (see Nonreturns, above).

For large populations, a 100% return random sample of 400 is usually sufficient for estimates within about 5% of population parameters. Then, if a return rate of 50% is anticipated from a mailed questionnaire and a 5% sampling error is desired, 800 should be sent. The disadvantage of this approach is that nonresponse bias is uncontrolled and may cause inaccurate results even though sampling error is somewhat controlled. The alternative is to reduce sample size (thus increasing sampling error) and use the resources thus saved for tracking down nonresponders. A

compromise may be the best solution in many cases.

While total sample size is an important question, returns from subgroups in the population also warrant careful consideration. If generalizations to subgroups are planned, it is necessary to obtain as many returns from each subgroup as required for the desired level of sampling error. If some subgroup is relatively rare in the population, it will be necessary to sample a much larger proportion of that subgroup in order to obtain the required number of returns.

Small populations require responses from substantial proportions of their membership to generate the same accuracy that a much smaller proportion will yield for a much larger population. For example, a random sample of 132 is required for a population of 200 to achieve the same accuracy that a random sample of 384 will provide for a population of one million. In cases such as the former, it usually makes more sense to poll the entire population than to sample.

References

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method.* New York: John Wiley.

Hinkle, D. E., Oliver, J. D., & Hinkle, C. A. (1985). How large should the sample be? Part II--the one-sample case. *Educational and Psychological Measurement*, 45, 271-280.