"A Guide To Family Intervention and Prevention Program Evaluation"

Glenda Kaufman Kantor and Kathy Kendall-Tackett (Editors)
University of New Hampshire

Edited and prepared for electronic dissemination by
Craig M. Allen
Iowa State University

December 2000

(Selected Sections)

Evaluation Designs

True Experimental Designs

True experimental designs compare people who have received an intervention ("treatment group") to an equivalent group who did not receive the intervention ("control group"). Subjects are randomly assigned to either the treatment or control groups; indeed, random assignment is the hallmark of the true experimental designs (a.k.a., randomized trials). In many circles, the randomized trial is the "gold standard" of quantitative research, reflecting its degree of methodological rigor.

"Rigor" refers to the degree to which evaluators can rule out alternate explanations for their findings. The true experiment, with its random assignment, allows evaluators to state with increased confidence that the intervention, and not some other factor, is responsible for the given results. What random assignment does (at least hypothetically), is to create equivalent groups, thereby controlling for the effects of other factors. Still, it will be necessary to monitor that equivalence to see if it is indeed true. This is the only design that allows researchers to specify cause-and-effect relationships (see Boruch, 1998; Fitz-Gibbons & Morris, 1987).

The true experiment, however, is not the only way to get at "truth." An example of this is in the relationship between cancer and cigarette smoking. For many years, cigarette companies hid behind the excuse that health officials could not "scientifically prove" that smoking causes cancer. In the framework of the true experiment, individuals would have to be assigned to "smoking" and "non-smoking" groups; obviously, this would have been quite unethical! Because researchers could not do this, it took multiple studies to account for all the other possible explanations (e.g., smokers may have a variety of other poor health practices that could account for higher cancer rates).

A similar problem arises in family violence research; researchers cannot assign children or adults to "abusive" and "non-abusive" families. In determining the relationship between family violence and other problems, evaluators must often account for other problems that might also be present in abusive families. In fact, a continuing controversy in child abuse research is whether the

difficulties that abused children experience are due to child abuse per se, or a general level of family dysfunction that often co-occurs in abusive families (Kendall-Tackett, Williams, & Finkelhor, 1993).

True experiments can be difficult to use in family violence research. The random assignment element can also make them difficult to use in an evaluation setting. Often, administrators and other program stakeholders that seek an evaluation want to know if their on-going program is effective. Random assignment would mean starting an evaluation with a set of clients who are just beginning the program: some would be assigned to participate, and others would be assigned to a control group. This might be practical for a program that is short in duration (such as a parenting class). It is also doable for a more intense program, illustrated by the following example. However, before attempting a randomized trial of a secondary prevention program, it is strongly suggested that evaluators consult with individuals who have expertise in this area.

Example 5.1: Evaluation Using a True Experimental Design

An evaluation of a program using a true experimental design was undertaken in New Jersey, involving over 200 families (Feldman, 1991). Families who participated in the study had been screened and referred for services by state and local agencies, and were randomly assigned to either the Family Preservation Services (FPS) program group (96 families) or a comparison group consisting of 87 families receiving traditional community services.

The researchers looked at a variety of family functioning variables as indicators of program impact. A comparison of these variables after the program showed that the FPS families had fewer children placed out of home, and also that these placements occurred at a slower rate (i.e. occurred several months later, on average, after services were received, than for the control families). Using random assignment enabled the researchers to attribute these differences between the two groups to the FPS program. This is especially important in this example because the life of any family is complex and involves many events and influences that might lead to a change in functioning, regardless of program participation.

Pros and Cons of True Experimental Designs

The randomized experiment is the most powerful research design for determining whether a program is effective. Random assignment is not always possible or practical, however.

- There is always the danger that the randomization procedure will not work, and that the groups are still actually different. This means that the program's impact is still unknown because scores might differ due to some other differences between the groups besides those that might be attributed to the program.
- When the program in question involves health-related or prevention services, it is problematic (and sometimes unethical) to deny services to people who need them just so that they can serve as a control group. If issues are not pressing, using a wait-list control group is an acceptable and often-used solution.

Bottom Line: True experiments are likely to be quite impractical for evaluating secondary prevention programs. They can be quite expensive and difficult to implement, should not be attempted without the assistance of an outside consultant. Further, there are often easier and less

expensive ways to gather the same information. However, a small-scale true experiment on a primary prevention program might be feasible.

Quasi-Experimental Designs

In many instances, random assignment is either not possible or not the most appropriate design to use. When that is the case, the evaluator can use a quasi-experimental design instead. Often, people are assigned to groups using a procedure called "matching." Clients receiving services are paired with people who are similar (e.g., by age, ethnicity, or marital status, etc.) but who are not receiving services. These designs are called "non-equivalent," referring to the lack of random assignment. Matching is the procedure used to create some equivalence. If people are in groups that cannot be broken up (e.g., a family), then matching can be done for groups rather than individuals. Some simple statistical analysis (e.g., Chi Square, t-test, Analysis of Variance, etc.) will help determine if sample groups are roughly equivalent, or if there are significant differences, before treatment even begins (see King, Morris, & Fitzgibbons, 1987; Mika, 1996; Reichardt & Mark, 1998).

Following are some examples of quasi-experimental designs that might be used:

Non-Equivalent Control Group Design with Pre- and Posttest.
Evaluations of new programs in multi-site agencies might be suited to this design. Sites not initially receiving the program, but which have the same primary focus and general location as the program sites, serve as controls. These control sites could be matched with program sites on several other characteristics that they have in common (e.g., stateside, near an urban area). The program sites should also be matched to control sits with families that are similar demographically. Care should be taken when using this approach, however. Families selected from different sites may differ in many ways that have nothing to do with the program. The most typical approach is to select some key demographic information (such as age range, education level, ethnicity, marital status, or number of children) to compare the two groups. Consultation with a statistician may help to determine whether the program and control groups are appropriate to compare.

An important point is that data must be collected in the same way for those receiving the program as for those in the comparison group.

Time-Series with Non-Equivalent Control Group. This design is identical to the one above, except it involves more than just one assessment at pretest and posttest. "Time series" simply means that data are collected at several (at least 3) time points before and after the program is implemented. The advantage to this approach is that the evaluator will get several "looks" at what is happening with each group before, after, and even during the program allowing for a greater understanding as to which changes are due to the program, and which changes are not.

Example 5.2: Non-Equivalent Control Group Design with Pre and Posttest

In a plan to evaluate the proposed United States Air Force FAP New Parent Support Program (NPSP), families would first be assessed for service needs and risk potential for family maltreatment. Then, each family would be identified as a "high needs" or a "low needs" family, based on several standard measures of family functioning. It would be important to design

measures for both. High-needs families would be offered more intensive home visitation services, whereas low-needs families would be eligible only for more standard educational and support services. Both high-needs and low-needs families would be re-assessed at a point after which most families have terminated services (one year) as to their current level of needs, stressors and family functioning.

The primary objective for the evaluation of the NPSP would be to determine whether the program reduces risk for abuse among high-needs families. Other comparisons could also be made between high- and low-risk families, and the relative impact of the respective services offered to each. The evaluation would allow an assessment of the stability of risk (needs) levels by looking at the patterns of changes in level of risk over time (see "time-series" designs for information on repeated assessments).

This is a good example of a quasi-experimental design in that families would not be assigned randomly to groups, but rather they would be assigned services on the basis of needs level. Having data both prior to the NPSP as well as after would allow for some conclusions to be drawn about the program effects. Families deemed high-needs who decline or do not attend services may be available to be part of yet another type of comparison group.

Example 5.3: Time-Series with Non-Equivalent Control Group

One large study by the Center on Child Abuse Prevention Research (Daro, Jones, McCurdy, George, Keeton, Downs, & Thelen, 1992) aimed to investigate the relative impact of 14 child abuse prevention programs in place in the greater Philadelphia area. It would have been completely impractical, and also prohibitively complicated, to recruit individuals and randomly assign them to programs. Instead the researchers tracked clients who were already receiving one of the programs, and compared outcomes.

The researchers were trying to measure rates of child abuse behaviors. This is a difficult and complicated task, as it occurs sporadically, and family members are reluctant to report behaviors. The best assessment strategy here was to use many assessment points over time, i.e., a time-series design. In the study, every client receiving services completed the Child Abuse Potential Inventory (CAPI). They also provided information on their level of satisfaction with services and demographic information at five separate one-week periods throughout the three-year evaluation. In this way, the evaluation families became accustomed to the assessments, and researchers got a "snapshot" of client characteristics (likelihood of abuse, demographics, participation patterns) at regular intervals.

Pros and Cons of Quasi-Experimental Designs

Pro: Quasi-experiments are close approximations of true experimental designs and can be used to investigate relationships between factors.

Pro: Quasi-experiments are the best option when random assignment is not practical, or does not make sense given the target population, resource constraints, research questions or ethical considerations.

Con: Because experimental and control groups are not formed by random assignment, there may be reasons other than the program interventions that explain differences between groups.

Bottom Line: Quasi-experimental designs are amenable to many types of prevention programs. Again, the evaluator must consider some of the resource considerations described with true experiments. "Big" quasi-experimental designs are going to be more expensive and more intense than "small" quasi-experimental designs (see below). They will also be easier to implement in primary prevention programs than secondary prevention programs. For assessment of secondary prevention programs, the assistance of an outside consultant is recommended.

Non-Experimental Designs

Non-experimental designs do not have comparison groups. These designs are used to assess program impact when there is no control or comparison group available, and usually involve time-series measurement. The soundness of these designs is affected by the number and timing of measurements. Non-experimental designs can be used for process evaluations when the primary purpose is to describe participants' experiences. One of the best ways to determine the process of change is to compare the groups performance or level of change to a baseline level established before the program began.

Types of Non-Experimental Designs

Single Group Pretest-Posttest Design. This design compares the same group of participants before and after the program. The purpose of the single group pretest-posttest design is to determine if participants improved after receiving the program. This design will not indicate, however, whether a program caused improvement in participants; there is no way to distinguish between changes over time due to other factors and effects specific to the program.

Single Group Time-Series Design. This design is similar to the time-series control group design described previously, except that no control group is used. It is said that subjects act as their own control group, in that comparisons are made between different time points for each person. This design is used to look at changes over time. The evaluator can collect data at regular intervals (e.g. daily, monthly, weekly) or at staggered time points (called "interrupted" time-series design).

Example 5.4: Non-Experiment Design for aTeen Pregnancy Prevention Community Outreach Program

The United States Air Force FAP Teen Pregnancy program uses infant and pregnancy simulators to communicate the costs of teen pregnancy and parenting. Included in the program is training in the following areas: self-esteem, values clarification, healthy relationships, intimacy without sex, and assertive communication. The goal is teen pregnancy prevention. It will be offered to youth group leaders and squadrons on air bases, and later in the schools.

This program could be evaluated using several scenarios. With a single-group posttest design, rates of teen pregnancy could be calculated for program recipients, and compared to prior year rates in the same school. Comparing pregnancy rates to those of same age teens in

demographically similar community schools would be an appropriate comparison.

Pros and Cons of Non-Experimental Designs

Pro: Single group designs are easier to implement, and less expensive than experimental, or quasi-experimental designs.

Pro: These designs can serve as pilots, and help identify important variables related to success in the program.

Con: Often participants are likely to improve over time without intervention of any kind, and these changes can be mistakenly attributed to the program under evaluation.

Con: Other events can change (e.g., a new community project is offered)

Bottom line: These designs are the easiest to implement, but the most difficult to interpret. While evaluators can avoid some of the issues involved in finding a matching comparison group, it may still be difficult to account for outside factors that may be responsible for the given results (e.g., the simple passage of time). However, for certain programs, a non-experimental design may be the only option. The evaluator may find this design helpful in the beginning phases of a project to assess needs within a community, or to get a general idea of what type of program might be helpful. Again, caution must be exercised in reporting and interpreting the findings, and in avoiding undue inferences about the "effects" of the program

Reviewing Design Options: Example for a Couples Communication Program

In the previous section, three basic types of evaluation design were described. More than one evaluation design can be used for the same program. Below, five alternatives to evaluating a program designed to increase communication skills in couples are presented.

1. One might use a **quasi-experimental design** with non-completers as the possible comparison group. This approach would be more cost effective than screening a large population for matched controls, and would also have the advantage of pre-collected baseline data from intake. With this approach a substantial proportion of non-completers could be identified, located, and interviewed by phone. Self-report measures on the same outcome instruments as those administered to completers would be obtained. Drop-out interviews would also provide a means to assess the reasons for program discontinuance or dissatisfaction.
2. One design modification would be to add on a new, time-two-only outcome measure for both experimental and comparison groups.
3. An alternative model would be to use a design similar to that described above comparing class or treatment group couples to matched couples receiving alternative programs.
4. Another evaluation could be conducted by using a **multiple comparison group** design with Group 1 (Communication Couples group), Group 2 (Alternate marital program), and Group 3 (Communication Group non-completers).
5. A **true experiment** could be organized with random assignment of families to either the Couples Communication program, or an alternate program, or a no "treatment" condition such as a wait-listed group or an attendance at one overview class only group. Long-term

follow-up would be desirable with this type of design. The strength of this design is its greater scientific validity over the designs noted above.

The information presented above is to assist the evaluator to select an appropriate design. Once a design has been determined, the scope of the evaluation being performed should be considered. This is the next topic.

Evaluation Scope

Impact evaluations come in all sizes. For any of the designs described above, it is possible to have versions that are "small," "medium," or "large." Factors that contribute to the "size," or scope, of the evaluation are described below. In particular sample and available resources are considered. Questions related to both of these factors are found on Worksheet 5.2.

Who Will Participate In the Study? Identifying The Sample.

In evaluating the program, all participating families can be assessed or a sample of clients who are "representative" of the target population can be drawn. Unless the number of participating clients is small, drawing a sample will be the most sensible choice. On Worksheet 5.2, questions are provided to assist the evaluator in thinking about the number of clients or families participating in the program under evaluation. One of the questions to be considered is the percentage of those participating that might yield an acceptable number for an evaluation. For example, if there are 9,000 parents who participated in a parent-education program, it might be decided that 10% (900) of the participants would yield an adequate sample size. Or if a program has multiple sites, the evaluator might decide to sample by site. For example, in the United States Air Force, if 30 air base sites are using a particular program, the evaluator might decide to sample clients from 20% (6) of the participating bases (Henry, 1998; Pecora, Fraser, Nelson, McCroskey & Meezan, 1995).

One factor that might influence the decision about the percentage of participating clients to be included is the size of the sample that each percentage yields. It is important to have enough subjects to demonstrate program impact, but not so many that the evaluation becomes unmanageable. A sample that is too small may lead to the false conclusion that an intervention has failed. The general rule is: the larger, the better. However, a large sample might not be possible, or even necessary. If resources are limited, it might be necessary to consult a statistician to know what the minimum necessary sample size would be for demonstrating an effect.

Technical Tip 5.1: Number of Subjects

It is important to have enough subjects in each group. Too few subjects will not yield enough data to show statistically significant effects even if these exist. A good rule is to have at least 50 subjects per group. The smaller the effect, the greater the number of subjects needed to detect it, and vice versa. Sample size considerations are particularly important when one is examining a low probability behavior such as the onset of physical aggression as an outcome indicator of program effectiveness in a non-high risk group.

Much has been written on the subject of sampling, and a full discussion of the various types of sampling is beyond the scope of this manual. However, there are two sampling methods that are easy to implement. One is the **simple random sample**. Using this procedure, an evaluator would randomly draw a sample of, say, 20% of the target population. To do this, one might randomly

select every fifth family from a list of all participant families. Or one might decide to collect data from every family that comes into the program in a given time period (e.g., every Tuesday and Thursday, or on alternate weeks).

A more sophisticated technique is the **stratified random sample**. This technique still involves random selection, but ensures that different groups are represented in the sample. Demographic characteristics are commonly used to create groups. For example, a sample may be stratified by age, or gender, or ethnicity, or some other important characteristic. If there is a small number from a certain group (e.g., single mothers), the evaluator might decide to "over-sample" this group (i.e., have a higher percentage of this particular group in the sample than is present in the target population) in order to have a sufficient number of participants for data analysis. The evaluator might decide to sample from both low and high-risk families in the proportions that represent them in the total sample. These sampling techniques can also be used to sample by site. In a stratified sample, sites can be selected that reflect the range of size, mission and location of all sites participating in the program (Pietrzak, Ramier, Renner, Ford, & Gilbert, 1990).

With any sampling technique, there must be a specified inclusion/exclusion criteria to minimize potential bias that can occur when a sample is drawn in a non-standard way (e.g., sampling only "compliant" or "nice" families). It is important to keep track of people who were asked to participate but refused. The number of those who participated divided by the total number of those who were approached allows the evaluator to calculate the compliance rate. For example, if 1,000 clients who were eligible to participate were approached, and 700 actually participated, the compliance rate would be 70%.

Are Sufficient Resources Available?

A second scope issue to consider is whether there are sufficient resources to conduct an evaluation. In Section 1 (Subsection "Get An Overview Of The Program"), lack of human or financial resources was listed as one of the potential pitfalls in an evaluation. A common mistake is to drastically under-estimate the amount of work and money necessary to conduct an evaluation. Below is a listing of some of the factors related to the amount of resources needed for an evaluation, and the adjustments that may be made in scaling down a project.

Project Scale

The more sites and/or subjects included in the evaluation, the greater both the complexity of data collection and expense of the project. If resources are limited, an evaluation of fewer sites and/or fewer subjects is the best choice.

Instrument Development

Developing data collection instruments for an evaluation can increase both cost and length of time necessary for an evaluation. If it is decided that developing instruments is necessary, they should be kept as simple as possible. (Developing efficient instruments is described further in Section 6, Subsection "Collecting New Data.") On the other hand, attempting to use existing data that is in poor shape (or unusable) can cost more in time and money in the long run.

Pretesting Instruments

An issue related to instrument development is pretesting. When developing an instrument, field testing will be needed. Pretesting can be brief or extensive. If resources

are limited, a less-intensive session of pretesting is warranted. Often, three to five pretests are sufficient.

Data Coding & Cleaning

Data coding and cleaning refer to preparing data for entry into the computer. The less coding and cleaning required, the faster (and cheaper) this step will be. Closed-ended response categories and concise questionnaires are the most efficient use of resources. Questionnaire design is described in detail in Section 6 (Subsection "Designing Survey Questions").

Bottom Line: Both design and scope influence the level of difficulty in implementation. For each design, there is a "big" and "small" version. Factors associated with increased expense/time are as follows:

multiple sites

multiple assessments

lengthy assessments

complex data that requires extensive coding and cleaning

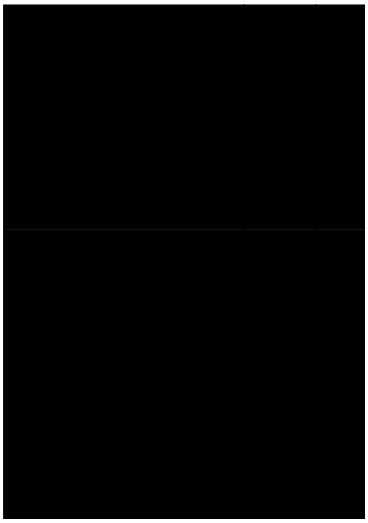
large numbers of subjects

If resources are limited, or results are needed quickly, a small evaluation with either focus groups or a survey is the most appropriate choice.

Elements of Design Selection	**Yes**	**No**
Will there be a comparison group?		
Will the comparison group be outside the program?		
Will group members be compared to themselves?		
Will group members be compared to those receiving services from an alternative but similar program?		
Will groups be divided based on amount of services received (i.e., measuring "dose" effects)?		
Will participants be randomly assigned to treatment vs. control groups?		
Timing of Assessments		
Will there be a pretest?		
Will there be a posttest?		
Will there be measures taken during the program?		
Will measures be taken at some specified point in the future?		
When?		

Sample Worksheet 5.2
Projecting the Scope of the Evaluation: USAF Example

Elements Related to Scope	Number/percent	Yes	No
Sample			
How many air bases will be used in the program?			
How many participants are estimated per air base site?			
Estimated total number of participants.			
What percentage of participants/air bases will be sampled?			
Total number of bases/participants in sample.			
Resources			
Anticipated number of subjects?			
Anticipated number of air base sites?			
Will it be necessary to develop new data collection instruments?			
Will data collection instruments require extensive pretesting?			
Will data need extensive cleaning or coding?			
Are there sufficient resources for evaluation:			
personnel?			
finances?			
time?			



Evaluation Methodology

This final portion of Section 5 focuses on the types of evaluation methodology used in collecting data. Either quantitative or qualitative techniques may be selected, or a combination of both. **Quantitative** analyses focus on testing hypotheses, and use structured designs and statistical methods to analyze data. This type of information needs standardization, precision, objectivity and reliability of measurement. Quantitative techniques are described more fully in Section 6.

Qualitative approaches, in contrast, gather data in a more open-ended fashion. Data collection usually occurs in natural settings, and focuses more on experiential or subjective aspects of a program. These data can include narrative accounts and may employ multiple data collection techniques (Worthen, Sanders & Fitzpatrick, 1997). Evaluators use qualitative designs to help them understand and describe program implementation, rather than to demonstrate statistically significant effects. With quantitative designs, the evaluator uses direct contact and experience with program families in order to gain in-depth and detailed understanding of the program. Qualitative designs use naturalistic methods of information-gathering (e.g., observation, interviews, case studies). These methods allow evaluators to explore experiences of the program for different individuals, in contrast to providing them an overall impression of its effects.

As Worthen and colleagues have indicated, "an unfortunate debate" (1997, p.71) has arisen about which of these approaches is "best." This debate is not productive. A combination of approaches frequently yields the most useful information. For example, when first approaching an evaluation, an open-ended approach can provide information that will assist the evaluator in developing more quantitative measures (such as a survey). Qualitative measures will provide illustrations and examples that stakeholders may find helpful in understanding the effect of the program on individuals. Further, stakeholders, and members of the general public tend to remember and be moved by case illustrations that describe an individual's experience.

On the other hand, data that are entirely qualitative can quickly become unwieldy. Processing this information is very labor intensive and may be well beyond the resources available. Quantitative data summarize information from large numbers of people in a concise way. Furthermore, the greater the methodological rigor, the greater the certainty that it is the program, and not some other factor, that accounts for the differences identified in family functioning. Rigor can be much more easily determined with quantitative methods.

Although a distinction has been drawn between these two methodological types, in the real world, these lines are often blurry. For example, in Section 6, an overview of survey research is presented. Many would consider this a "quantitative" technique. However, if the questions on the survey are open-ended rather than closed-ended, they are really more qualitative. Further, questionnaires quite often contain both types of questions. Moreover, focus groups, usually considered a "qualitative" approach, frequently yield information that can be coded and quantified.

Below, three qualitative techniques are described that are easy to use and valuable at all stages of program evaluation. These techniques are also handy because they allow the evaluator to "take a quick pulse" of the program, and make decisions about how it is functioning.

Types of Qualitative Approaches

Open-Ended Interviews

Rather than using closed-ended questions, open-ended interviews have several advantages:

Clarity: The interviewer can clarify questions, and avoid a problem such as illiteracy or language barrier.

Richness: The interviewer can pose more complex questions and make observations about the respondent's appearance and behavior.

Completeness: The interviewer can minimize missing and inappropriate responses.

Control: The interviewer can control the order of questions.

Strategies For Involving Respondents

Those planning to conduct open-ended interviews should investigate various factors that will help to increase the response rate, such as the best time of day to call or visit. Before conducting in-person interviews, an advance letter could be sent to the sample respondents, explaining the purpose of the survey and that an interviewer will be calling soon. Respondents must be informed of the voluntary nature of the interview and how the answers are to be used before they agree to participate.

Interviewer Training

Interviewers must be carefully trained in the techniques of interviewing such as making initial contacts, listening skills, and avoiding influencing or biasing responses. It is possible for an inexperienced or improperly trained interviewer to skip certain questions that he or she is uncomfortable with, or to not listen carefully to respondents. Practice interviews are necessary before the interviewer conducts any interviews. The interviewer needs to be fully familiar with each interview question, to be able to assist respondents who might need questions clarified. The interviewer also needs to be well-versed in the ethics of interview research. This includes maintaining confidentiality and not coercing or pressuring participants to participate, or answer specific questions.

Focus Groups

Focus groups are very much like individual interviews except that they are done with a group of people. They involve discussion of a topic that is the "focus" of the conversation, and typically last 1 ½ to 2 ½ hours.

Purpose

The most common purpose of a focus group is to help the evaluator learn more about a little-known topic. It is especially helpful for providing information about how people think or feel about products or services, and is frequently used in marketing research. This same technique can be used to help the evaluator to learn more about the needs of a particular population, or how these individuals reacted to a prevention program. In addition, focus groups are useful in conjunction with a quantitative study in explanations of results. One should conduct 3 to 12 focus groups on a topic, because various combinations of participants will yield different overall responses to questions and issues.

Selecting the Participants

Focus groups usually contain 4 to 12 members to allow ample opportunity for expression, and to enable the leader to reasonably manage and document the discussion. Ideally, the participants should regard each other as equals so that they feel free to speak what is on their minds. However, depending on the resources available it may not be possible to constitute groups in a singular manner (e.g., low-income fathers). Often groups are selected to get opinions from a cross-section of people (e.g., active duty military mothers and non-military mothers). Unlike regular survey research, random samples are not a priority. In terms of group composition, group members are selected who are knowledgeable about or have experience with a particular topic. It is not necessary to limit focus groups to program clients. Informants can be used to problem solve with staff members, or to assess a need in a specific community. Informants should be selected on the basis of their knowledge and experience with the topic being explored. Another factor to consider is the impact of group members on each other. For example, women are often less talkative in groups with men than they are in groups that are all women.

Technical Tip 5.2: Helps for Successful Focus Groups

Identify potential focus group members well in advance of the planned meeting day.

Send advance letters or make phone calls as the day approaches.

Offer a small incentive in advance.

Supply beverages and snacks.

Break the ice with small talk first.

Explain the goals of the session.

Make name badges for everyone and have participants place them where everyone can read them.

Setting the Ground Rules

The facilitator's job is to keep the discussion focused, and allow everyone to express their opinions. Focus group facilitators must understand the cultures of the group members, and must make sure that each participant has a chance to speak. Moreover, they must let participants know that there are no right or wrong answers. All comments, both positive and negative, should be welcome. Also let participants know that they do not have to feel "on the spot" to answer something that they are uncomfortable about (one of their rights as research participants). Ground rules should include telling participants that they may simply say "pass" if asked a question they do not want to answer. Ground rules should also include getting consent to tape if this is part of the plan, and explaining reasons for taping or taking notes. Assurances of confidentiality are important and should include telling group members that no names are attached to individual comments, and that tapes will be destroyed after they are transcribed.

The Questions

The interviewer/facilitator uses pre-written interview questions developed as part of an interview guide, asks the group to respond, and helps to promote dialogue among the group. The interview guide is usually no more than 12 to 15 questions. Note that in some cases it may not be possible to fully cover more than a few questions, but it's nice to have some extras, especially if some questions "fall flat." The facilitator must be able to listen carefully and use appropriate non-leading probes to elicit more information from participants.

Taping and Analyzing the Focus Groups

Focus group sessions are usually tape recorded so that the discussion can be transcribed after the session when it is convenient. The analysis of focus group data can cover a wide range of specificity. In some cases, all that is needed is a brief summary of what was discussed in the group. For more rigorous approaches, the first step is transcription of the group tape. The tape should be transcribed verbatim. The evaluator then examines the written transcribed comments for common themes and insights, within and between groups. Direct quotations can be used as part of the presentation of findings.

Example 5.5: Focus Group Structure in the USAF Family Advocacy Program Evaluability Assessment

Each of the field staff focus groups conducted by the UNH team during their evaluability assessment began with a question on the most challenging aspect of FAP for individual members. The format for each group was similar although groups were allowed a lot of latitude to discuss issues that were salient to them. Facilitators encouraged participants to raise issues as they arose during the course of the discussion instead of enforcing a more rigid structure. Consequently, there were some distinct aspects to the content and direction of each group. Initial questions are provided below:

What do you feel is the biggest challenge at your air base in relation to the program you are

involved in?

How are families screened for different programs?

What kind of risk assessment is used to determine who participates in the programs?

What is working well in your program?

In formal content analysis, coders search for themes in the transcripts. Each transcript should be independently reviewed by two different analysts, who then compare their reviews for content and consistency. They must agree on the content and meaning of the transcripts and notes. Analysts look for themes, key terms and concepts, and contrasts in the data. Points that are not clear are then checked, and verified by reviewing tapes and notes, and through discussion with the facilitators. Coders need extensive training on how to code focus group (or other qualitative) data. Formal content analysis can be very labor intensive. This process can be helped by computer programs designed to follow "data-making" rules established by the investigator, but can still be demanding.

Evaluators should not rely solely on this type of data. The lack of representativeness in the samples can be misleading. On the other hand, focus groups can highlight issues that evaluators may have missed and can add richness to quantitative data that are collected.

Direct Observation

Using direct observation for assessing families' interaction has advantages, especially when trying to determine whether child maltreatment is occurring. Direct observation of interactions between children and parents is useful for assessing the basic elements of a child's health, safety, and physical development, while eliminating possible biases from the children's caretakers. Direct observations may be conducted in the home, or the clinic, with each setting providing its own type of information.

Observations in a clinic-type setting are useful when observing behavior between individuals, such as between parents and children. Behavior in the office is often assumed to be a client's "best." Abusive behavior observed in the clinic (when families know that they are being observed) often leads staff to conclude that the situation is worse in private.

Home observations are useful for understanding the whole scope of influences on a family. Home visits may help pinpoint problems quickly in a way that traditional interviews cannot. For example, a breastfeeding baby with slow weight gain may be distracted from nursing at home because of the chaotic or noisy environment.

Pros and Cons of the qualitative measurement technique

Pro: Qualitative techniques work well for exploratory research where pre-determined response categories may not need to be anticipated or developed (see also Section 6).

Pro: Qualitative techniques yield in-depth, detailed, and rich information.

Con: Analysis of qualitative data can be difficult and time-consuming because responses are neither systematic nor standardized.

Con: Qualitative data collection is often more labor-intensive; it requires more time and effort to interview or observe than to administer questionnaires to a large group.

Bottom Line for Qualitative Methods: The qualitative techniques that have been described above will be helpful during many phases of the evaluation. These techniques, and the survey techniques described in Section 6, will yield the most comprehensive data. (In Section 6, further information is provided to guide the evaluator through an impact evaluation. Both data coding from existing records and designing surveys are

Designing Data Collection Instruments

Key Points & Introduction

Key Points

Data can be collected from existing records or by developing new data collection instruments.

Information can be gathered from existing records by developing a coding sheet or by coding it directly into the computer. Decisions about the logistics of this are based on client confidentiality, availability of the records, and availability of staff who can code these records.

A questionnaire can be developed to collect new information.

Questionnaires can be self-administered or administered by an interviewer in-person or by telephone.

Survey questions can be open-ended or closed-ended. Order can influence the answers respondents give.

There are many tools an evaluator can use for data collection. The evaluator's choice of tools depends on the type of information needed, the amount of time available for data collection, and available resources. For example, when the goal is to determine "client satisfaction" for program supervisors), interviews and focus groups may be the best method for gathering the desired information. On the other hand, if a major stakeholder wants information on the efficacy of the program, it may be necessary to draw a representative sample of clients and present information that was collected in a standardized way.

With regard to data collection, the evaluator should to remember that "less is more." It is important to be specific about the type of information needed (Section 1, Subsection, "Determine The Audience For The Evaluation"); the evaluator should try not to succumb to the temptation to collect more data than is actually needed. Chances are that the cooperation of field personnel will be required either in gathering existing data or collecting new data. If the job is either too big or perceived as a waste of time, it is unlikely that field personnel will cooperate in this process.

Using Existing Data

The use of existing data is often helpful, and can save considerable time and money if the information is of good quality. The evaluation may call for data from records such as class attendance, lists of learning objectives, pre- and post-tests, prevention files and scores on standardized instruments. The disadvantage of existing data is that the evaluator has no control

over which data are collected, the quality of the data, and the methods of data entry, filing, and storage.

When the evaluator has decided to pull information from existing forms, it will be important to plan carefully and be specific about what type of information is needed. For example, a form may have information about gender of the parent and the child, age of the child, employment status, and whether the family receives ancillary support from organizations like WIC. The evaluator will need a systematic way to transform this information into numerical data.

Type of Variables

The evaluator can enter this information directly into a spreadsheet (this is discussed below). Alternatively, it might be more practical to record this information on paper and enter it into the computer at a later date. Below is a sample coding form that could be used to extract information from existing records. Notice that each response category (e.g. male/female) has a corresponding numerical value. Information that already has a numerical value (e.g., age) does not require an additional numerical value; the variable name can simply be written and a space left for the coder to enter this information.

Example 6.1: Coding Demographic Information

Gender of child
male…….1
female……..2

Age of child
_____

WIC recipient
yes……1
no…….….0

For many variables, the numbers, as numbers, are arbitrary and meaningless. For example, by convention, "male" is typically coded "1," and "female" is coded "2" (or female can be coded as "1" if preferred). But if an average for the variable "gender" were computed, the number derived would make no sense. These types of variables are called "categorical." Categorical variables are useful in describing the sample. For example, the evaluator may want to know how many male and female children are in a sample. A frequency distribution with the number of "1's" and "2's" would provide this information.

Technical Tip 6.1: Coding Yes/No Variables

At first glance, the codes for yes/no questions may also appear arbitrary. However, a general rule is to code 'yes" as "1," and "no" as "0." This is helpful when deciding to add together a series of yes/no questions (such as number of agencies that are involved with a particular family, or the number of life stressors that a client reports). If yes/no has been coded in this way, the sum of different variables in a set is meaningful. On the other hand, if "no" is coded as "2" (as many researchers do), the evaluator will have more difficulty interpreting the sum of various yes/no questions. Or the evaluator may have to "recode" the no's in the statistical program, adding another step to the process.

Coding Forms

Should the evaluator develop a coding form or simply enter the data into the computer? That depends on the resources available for the evaluation. It will need to be determined if the data is to be entered into the computer at each site or sent to a central office. If sites have the facilities to enter data, it might make more sense to enter this information directly into the computer rather than completing a coding sheet. This may be particularly helpful if the statistical package being used presents a data template so that coders do not have look up codes every time they enter information. On the other hand, some statistical packages do not have this option and it may make more sense to have the coder simply circle the appropriate code on a sheet. Further, some facilities or sites will not allow the evaluator to remove files from a specified area. Unless the coder has a laptop computer or access to a computer with appropriate software, a coding sheet may still be the best bet.

If a pre-coded coding sheet is developed, it should be pretested on any and all forms being used. Following in-house testing, the pre-coded coding sheet should be sent to a few other sites for them to try. This type of pre-testing will be an efficient time save in the long-run and will help minimize confusion about how this information should be coded.

Technical Tip 6.2: Developing a Coding Form

Sometimes when individuals develop coding forms, they are tempted to squeeze as many questions as they can on a page. As thrifty as this may be, questions instead should be spread out on the page. Tightly spaced questions increase the number of coding errors. Instead, codes should be listed, with ample space left between questions. The evaluator should use only one column of information on a page, not two. This format makes it easier to both code and enter data, and will save considerable time.

The "Coder" of Data

Who should code the data? Again, this depends on the type of data and resources available. In many cases, someone at the site can code data out of the files (although it might be necessary to give release time or compensate the coder for extra time). Sites may decide to hire an extra person to help with this task. Or the evaluator may decide to pay someone to photocopy the relevant documents and have them coded at a central location. The types of data that are the most amenable to coding are unambiguous variables such as gender, age, or duty status. When the coder must make a judgment call, more opportunity for coding error occurs. Although qualitative or open-ended coding can be done, it requires more training of coders and introduces greater chances for error (Sudman & Bradburn, 1982). For variables where coders must make judgment calls (such as assigning a single code to a paragraph, or page, of narrative data), the evaluator should provide coding guidelines that will help coders make these decisions. Pre-testing the coding form will help the evaluator to anticipate the types of questions and concerns that might arise. Last, interrater reliability assessments could be conducted to help the evaluator to determine the extent to which two or more coders are in agreement in their judgements. Interrater reliability assessments help to determine if coders are interpreting data in consistent fashion.

Client Confidentiality

A final concern has to do with protecting the confidentiality of client responses (Fowler, 1998). Much of the information contained in client records is highly confidential and must be protected.

One way to protect confidentiality is to identify families only by number. This number can be assigned by the evaluator or taken from existing records, such as the last four digits of the Social Security number. Another important issue is how to track families through multiple data collection opportunities. The evaluator may need to develop a protocol about how a number should be assigned so that data from multiple sources is matched to the right family. The evaluator may also decide to keep a master list of family names and their numbers in order to keep track of information relevant to each family. Obviously, this list must be handled carefully and kept secured (e.g., in a locked filing cabinet) when not in use. Below is an example of a study in the child abuse literature that drew all of its information from existing sources

Example 6.2: Use of Existing Documents

Use of existing documents can be a rich source of data. One example is the *Maltreatment and School Achievement* study conducted at the Family Life Development Center, Cornell University (Eckenrode, Laird & Doris, 1993).

Records that were coded included information from the child abuse register and school records. The register provided information on the type and severity of maltreatment and the demographic characteristics of the child and family. School records provided information on academic performance, grade repetitions, school transfers, home moves, and disciplinary actions.

If there are concerns about confidentiality, it may be more appropriate to have the data coded by someone else. The evaluator may decide to have someone already familiar with the families do the coding. Such an insider might be more sensitive to the family's issues and, therefore, have a greater appreciation for confidentiality. On the other hand, by using someone from the outside who is not familiar with the families, confidentiality might not be an issue at all. Yet another option is to have someone obliterate the names and other identifying information before turning the information over to a coder. Again, these types of decisions need to be made on a case-by-case basis. For some small sites, it might make more sense to have the information coded on site. For sites involving multiple families, the information might be better handled at central coding location.

Items vs. Scale Scores

Another type of existing information is scores on previously administered standard instruments. With standard instruments, the evaluator can simply enter a summary score, or enter all the individual items. Entering a final score is obviously faster (and may be the only information available). On the other hand, entering the responses to each item provides for more flexibility, and allows the evaluator to look at individual items or sub-scales.

As can be seen, existing files can provide a rich source of data. However, it may also be that information is still needed that is not in the files. How to collect this additional information is the focus of the next section.

Collecting New Data

Once the type of information still needed has been determined, a measure that will collect it must be located or developed. In this section, the technique most likely to be used, survey research, is described. A survey is simply a series of questions. Researchers give surveys to individuals who

answer the questions and then return the survey to the evaluator. The evaluator then analyzes the responses to the survey questions, and hopefully draws meaningful conclusions from those answers about the program under evaluation (Fowler, 1998; Mangione, 1998).

The basic survey design is a skill that takes time to master. As the evaluator approaches the task of developing a new questionnaire, allow sufficient time to pre-test items. Learning what others have done can be very helpful. Using questions developed by others can speed the process considerably, even if the evaluator has to modify them (Sudman & Bradburn, 1982). Sometimes a pre-existing measure can be used (e.g., a depression inventory) as part of the survey, and then filled in with other items the evaluator has developed. Some of the basics of good survey questions are described below.

Designing Survey Questions

The evaluator may decide to write the survey questions when an appropriate data collection tool cannot be found (Sudman & Bradburn, 1982). Good questions are those that:

ask for specific information

address only one topic per question

are not too general or vague

do not contain hidden biases that might influence the respondent's answer in any way

are necessary to ask, not just interesting to know about

are clearly worded so that they cannot be misinterpreted.

Some questions are written to ask about respondents' opinions and attitudes ("How did you feel about…"?), while other questions are concerned with facts or behaviors (people's health, income, housing). Below are three basic types of survey questions (Fowler, 1998).

Demographic/Other Content

Surveys frequently include questions that ask about the respondents themselves (demographic information). These questions help the survey researcher classify their responses (such as age, gender, marital status, occupation, and place of residence).

Opinions and Attitudes

The evaluator may want to ask respondents what their opinion is on an issue, how strongly they feel and why, their interest in the issue or other questions that try to get at a person's feelings. For example "How did you feel about having a visiting nurse come to your home"? "Do you feel that the program was helpful to you"? There are no right or wrong answers to these types of questions.

Knowledge

These questions assess respondents' knowledge on a particular topic, such as knowledge of appropriate developmental milestones. Unlike opinion questions, there are right and wrong answers.

Behavior

Questions about behavior ask respondents to indicate whether or not they have participated in a certain activity, and how often they have done it. These types of questions are important for assessing potential or actual family violence, for example.

However, respondents may under- or over-report their behaviors because they really do not remember, or because they want to appear favorable to the interviewer or evaluator.

Question Format

In the previous section, information about the content of questions was provided. It will now be important to describe how these questions can be formatted. Survey questions may be open-ended or closed-ended (forced-choice). Most questionnaires will probably contain both types of questions. The following examples are taken from a form developed by the USAF Family Advocacy Program.

Open-ended Questions

Open-ended questions allow the respondent to answer questions in an unstructured format; there are no pre-coded response options. Open-ended questions are typically structured as a question followed by blank space for the respondent to write out an answer. Similarly, in a spoken interview, the interviewer asks a question, the respondent answers, and the interviewer records the response. An example is listed below.

What was most helpful among the services you received from your FAN (Family Advocacy Nurse)?

_____

_____

One of the uses of the open-ended format is in pre-testing and questionnaire development. Open-ended questionnaires allow the evaluator to collect information that can be used to develop closed-ended categories for a future assessment. Answers, or themes, that occur frequently are good candidates for closed-ended response categories (Sudman & Bradburn, 1982).

Open-ended question formats have other advantages and disadvantages:

Advantages

The respondent uses his/her own words to answer the question, without being confined to narrow options provided on the survey.

They give the evaluator an idea of exactly how the respondent feels or thinks.

They provide specific information to be used to illustrate overall themes among many respondents.

Disadvantages

They do not provide information that is easily summarized into numerical data.

They are more difficult and time-consuming to analyze.

In a self-administered format, the responses depend on the reading and writing ability of the respondent.

They can be more time-consuming for the respondent to answer.

Closed-ended Questions

Closed-ended questions come with a set of pre-coded response options. There are several types of closed-ended responses. These include rating scales, true/false or yes/no questions, or

questions where respondents must circle the appropriate code. The following are examples of closed-ended questions.

Rating Scales

Questions that use a rating-scale allow the respondent to choose only one option. Typically, this is a 3, 5 or 7-point scale. A question or statement is presented, and the respondent is asked to circle a value on a numerical scale that corresponds with the best answer, such as in the following question (Fowler, 1998).

> *On a scale of 1 to 7, how would you describe your labor and delivery, with "1" being "very easy" and "7" being "very difficult"? (please circle)*
>
> 7 6 5 4 3 2 1

Check the Answer Questions

These questions present the respondent with a question or statement, and then a set of possible responses. The respondent simply places a check-mark next to the response (or responses) that apply.

How old are you? (Please check one)
15 to 19____
20 to 29____
30 to 39_____
40 to 49_____
50 to 59_____
60 or older_____

Technical Tip 6.3: Number Categories

It is preferable to have numbers previously assigned to each category (e.g., where the "15-19" category is coded as "1"). The reason this approach is preferred is because it saves time when this information is entered into the computer. Age information can also be gathered using a fill-in-the-blank format since it is already numerical data.

Did you experience any of the following complications with your delivery or afterwards? (Circle all that apply)

	Yes	No
Problems with the "after birth"	1	0
Hemorrhage	1	0
Seizures	1	0
Breech baby	1	0
Other (Specify) _____	1	0

Technical Tip 6.4: "Other" Categories

It is a good idea to include an "other" category in case all possibilities are not covered in the response categories. This way, the respondent can write in an answer that is not covered.

Fill-in-the-blank Questions

Some questions ask the respondent to fill in an answer:

In all, about how many years have you and your spouse/partner lived together?

years_____

True or False Questions

For the following statements, the respondent is asked to circle either T (True) or F (False).

Please read the following statements and circle true or false.	***T F***
There is a feeling of togetherness in our family.	*1 0*
There is plenty of time and attention for everyone in our family.	*1 0*

Yes or No Questions

For the following items, the respondent is asked to circle either "Yes" or "No."

Did your FAN (Family Advocacy Nurse) refer you to any of the listed services? (Please answer each question by circling Yes or No.)

	Yes	***No***
Parenting class services	*1*	*0*
Mental Health	*1*	*0*
Financial	*1*	*0*
Stress Management	*1*	*0*

Closed-ended questions have other advantages and disadvantages in comparison to open-ended questions (Sudman & Bradburn, 1982):

Advantages

Closed-ended questions help the respondent to answer the questions more easily.

They help the evaluator to collect and summarize responses more efficiently.

Disadvantages

Closed-ended responses limit the respondents' answers.

They inhibit the researcher's ability to understand what the respondent really means.

Putting Questions Together

Once the questions have been designed, the next step is to put them together in a survey. The evaluator can weave together newly developed questions, questions from other surveys, and questions from standardized instruments. Guidelines about how to put a survey together are provided below, including a rationale for why question order has the potential to influence respondents' answers.

Technical Tip 6.5: Using Questions from Other Forms

If to save time, the evaluator decides to use questions from other forms, be sure to ask permission from the author of the original form. The evaluator may have to modify the question to make it more appropriate to the survey (e.g., the survey may need to be changed from a self-administered form to one read by an interviewer). Even in its modified form, be sure to reference the original source in any reports from the evaluation (Sudman & Bradburn, 1982). Resist the temptation to use all standardized instruments in the survey. This can drastically increase the length of the questionnaire and may not add a worthwhile amount of additional data.

Question Order and Transitions

It is important to consider the order in which questions are placed on the survey. Question order can affect the responses to survey questions. The first few questions that are asked can set the tone for the survey. If respondents become bored or feel threatened by sensitive questions early in the survey, they may not complete it, or they may not answer the questions seriously. The answer a respondent gives to a question on a survey or interview can affect the way the respondent then answers the next questions. In this section, general guidelines on question order are provided.

Surveys begin with an introduction. Below is an example of an introduction to an interview, but something similar could be written at the beginning of a self-administered questionnaire.

> *"In this survey, I will be asking you some questions about what's been happening in your life since the baby was born. Before we begin, I want to remind you that all your answers are confidential and you can refuse to answer any question that makes you uncomfortable or that you do not want to answer. You can also stop the interview at any time. Do you have any questions?"*

The evaluator might begin with some basic demographic questions such as age, marital status, number and ages of children, and duty status. The most accepted strategy is to begin the survey with the least sensitive questions, and gradually build to the most sensitive questions. Similar items should be grouped together. Transitions between sections help the respondent to move into the different parts of the survey. For example:

> *"Now I want to ask you some questions about some stressors in your life. Please answer yes or no to let me know if any of these things have happened to you in the past six months."*

Asking questions about sensitive topics is also an art. There are ways to do it that gather the desired information, and yet are respectful of the respondent's feelings. Suppose the evaluator wants to know if the respondent has a history of childhood sexual abuse. The questions need to start gently, while reminding respondents that they have a right to refuse to answer. This type of

approach works well with many kinds of sensitive topics and can be applied to both interviews and self-administered forms.

> *"In this section, I'm going to ask you some questions about some things that may have happened to you in childhood. Before we start, I'd like to remind you that you can refuse to answer any question and all of your answers are kept strictly confidential."*

A yes/no format, where a list of items (such as a list of sexual acts) are read, and respondents then asked to indicate "yes" if this ever happened to them, works well and seems to be less invasive than to ask them to recount their experience in an open-ended fashion. This can also work well for a listing of other highly sensitive topics such as criminal activities or drug use.

Household income is another sensitive subject. For instance, asking for an income range rather than asking respondents to openly state their income is a more sensitive method of gathering information about income. Respondents interviewed in-person should be handed a card and asked to pick a letter that corresponds with the range of their income level. In a telephone interview, the list of income categories is read and respondents are asked to pick one. Below is a sample question that asks about income. The income ranges are adjusted to meet the needs of the evaluation.

> *What is the total income in 1999 for you and the members of this household who are related to you, before taxes and other deductions. Just give me the letter.*
>
> *A. $15,000 or less*
> *B. $15,000 to 17,500*
> *C. $17,500 to 20,000*
> *D. $20,000 to 25,000*
> *E. $25,000 to 30,000*
> *F. $30,000 to 35,000*
> *G. $35,000 to 45,000*
> *H. $45,000 to 60,000*
> *I. $60,000 or more*

The evaluator will also need to be aware that question order can affect answers respondents give (Schuman & Presser, 1981). Question-order effects seem especially likely if asking questions about sensitive topics. The following is an example of how question order influenced responses to two questions about partner violence:

Example 6.3: Question-Order Influence

Moore and Straus (1998) studied whether the order in which two questions dealing with approval of slapping a spouse were asked in a telephone survey would affect respondents' answers to those questions. The two questions were:

Are there any situations that you can imagine in which you would approve of a husband slapping his wife's face? (question A)

Are there any situations that you can imagine in which you would approve of a wife slapping her husband's face?(question B)

Respondents answered either "yes" or "no" to each.

Moore and Straus (1998) presented half of their telephone survey sample with question A first,

and half with question B first. They found the following results:

Question A (Husband slapping wife) presented first

10% of respondents said they approved of a husband slapping his wife
22% of respondents said they approved of a wife slapping her husband.

Question B (Wife slapping husband) presented first

19% of respondents said they approved of a husband slapping his wife
43% of respondents said they approved of a wife slapping her husband

When Question B was presented to respondents first, the percentage of those who approved of slapping in both questions almost doubled. Moore and Straus explain this effect by noting that there seems to be a double standard that says it is permissible for a wife to slap her husband, but not for a husband to slap his wife (Straus, Kaufman Kantor, & Moore, 1994). There is also a tendency for respondents to want to appear fair in their answers. Therefore, when respondents are asked first if it is acceptable in some circumstances for a wife to slap her husband (the more acceptable according to the double standard), more are likely to answer "yes". Those who do will also tend to answer "yes" to the question about husbands slapping their wives in order to remain fair in their answers. No matter what the explanation for this finding, it is clear that the order of the slapping questions in Moore & Straus's (1998) study did affect respondents' answers (Moore & Straus, 1998).

Pretesting

Pretesting is a way to gauge the appropriateness of questions or the survey as a whole (Suen & Ary, 1989). Before any formal pretest, the questions aloud should be read aloud to a colleague. Many errors can be spotted this way, including typos, awkward wording, and response categories that make no sense. The data collection tool used must be appropriate to the age, education level, and cultural background of the participant. If multiple sites are participating in the evaluation, it is important to have a pretest group that is similar to all evaluation subjects. Staff at other sites (especially those with an atypical group of clients) could also pretest the questionnaire. To make certain that the tool is appropriate, a small sample of participants (6-10) who are similar to the target population should be selected and administered the survey. Their responses can then be examined for possible problems with question order, wording, and and their feedback on the questions gathered. Pilot test participants can also be asked if the questions were clear, and the response format made sense to them. In addition, pilot testing allows the researcher to test the survey administration procedures (time limit, type of location, instructions for completing). The evaluator may then change the questions/survey according to pilot participants' feedback and suggestions.

This data should not be used as part of the evaluation study per se; the purpose of the pilot test is to check the measurement instruments and data collection procedures before any actual evaluation data are collected.

Survey Format

Another factor to keep in mind is the format of the survey. Are the questionnaires going to be self-administered or will they be administered by an interviewer? Interviews can take place either in person or by telephone. Interviews tend to be more expensive to administer than mail

surveys, but compliance rates (percentage of people who agree to participate) are generally better, with less missing data. Self-administered questionnaires are substantially less labor intensive to administer (questionnaires are simply handed or mailed to respondents), but there may be difficulties if some respondents find the questionnaire difficult to understand or confusing or if they do not read at the level of the questionnaire. Further, self-administered questionnaires are easier to "turn down" than a questionnaire with a real person at the other end. These issues are discussed below.

Self-Administered Questionnaires

Self-administered surveys can be administered in-person (such as when clients come into the clinic) or via mail. Mail surveys can be relatively low in cost, and may be the technique most likely to be used. As with any other survey, problems exist in their use when insufficient attention is given to getting high levels of cooperation. If only a small percentage of the families approached agree to complete a form, data may be biased in that only the more compliant respondents participated. This group may differ in some unknown way from the group who refused to participate (Schuman & Presser, 1981). The closer the compliance rate is to 100%, the better the quality of the information. A minimum of 60 to 70% compliance is typically acceptable (Fink & Kosecoff, 1985).

Increasing Compliance

There are several ways to increase respondent cooperation. First, respondents can be contacted before they are mailed the questionnaires, either by letter or telephone, to let them know a survey is on the way. The form itself should arrive with a cover letter, and a self-addressed stamped envelope for returning the completed form.

The cover letter should describe the purpose of the study and explain that all answers are confidential. There should be a phone number that respondents can call with questions, and instructions about what to do with the completed form. If the respondent will receive any type of compensation or token of appreciation, this should be described as well (Mangione, 1998).

The questionnaire should be as concise as possible. The practice of grouping multiple items together in order to save paper should be avoided if at all possible. If the form looks like it will be difficult to complete, the respondents are less likely to participate. To the extent that it is possible, try to have the form pre-coded since this format is faster and easier for the coder, and respondents have only to circle or check a response rather than writing out an answer (Sudman & Bradburn, 1982). During pre-testing of the form, the length of time it will take a respondent to complete the form can be determined. Trim as necessary.

Good follow-up will also increase compliance. After a pre-determined amount of time has elapsed since mailing the questionnaires (e.g., 3 to 4 weeks), reminder post cards or letters can be sent to respondents who have not yet returned their surveys. Also, additional survey packets can be sent (even though this increases the cost). This can be repeated two or three times, with the compliance rate increasing with each mailing. Follow-up telephone calls can also be used, if feasible (Mangione, 1998). Below is an example of a USAF program that was evaluated via a mail survey.

Example 6.4: USAF Mail Survey

Bowen (1984) used a mail survey to evaluate the US Air Force Family Support Center Program.

The evaluation team sent questionnaires to Air Force members, civilian spouses, and adolescents. The survey was publicized in the base newspaper before the survey was conducted. Then, each member of the sample was mailed a survey packet. One week after the first mailing, a postcard was sent to each respondent. The purpose of the postcard was to thank those who had responded, and as a reminder to those who had not yet responded.After two weeks, Bowen sent an additional cover letter and a survey to non-respondents, or used follow-up telephone calls to remind individuals to complete and return their surveys. Three weeks after the first mailing, Bowen sent a reminder letter and a fresh questionnaire to non-respondents. Finally, two weeks later, a final reminder letter and a fresh replacement questionnaire was sent to all non-respondents. All reminder mailings included postage-paid self-addressed envelopes and appropriate cover letters reminding respondents of the importance of the survey and asking them to please respond. Bowen's response rate was high (Bowen, 1984).

In reviewing the above example, one might wonder whether all the additional effort of Bowen was excessive. Nevertheless, the decision about the amount of follow-up of respondents is an important one that will need to be made. In almost all cases the compliance rate is more important than raw numbers. The information gathered is likely to be more accurate if there is a higher compliance rate. For example, if the evaluator were to send out 1,000 questionnaires with 700 returned (compliance rate=70%), this would be preferable to 700 questionnaires returned out of 2,000 (compliance rate=35%). The money saved by not sending out the additional 1,000 questionnaires could be used to send reminders.

Bottom Line: If financial resources are limited, the best strategy is to send fewer questionnaires with good follow up (increasing the compliance rate).

Self-administered questionnaires can also be completed in person. Questionnaires can be distributed to clients when they come into the office for appointments, or when a home visitor sees a client in his or her own home. The compliance rate for these types of questionnaires is likely to be higher than when the questionnaire is mailed, but the form will still have to be concise so that it can be completed in a brief amount of time. The main concern is distributing the questionnaires in a non-biased way. The evaluator will need to develop a protocol for distributing questionnaires and tracking refusals. For example, the evaluator may decide that all clients who come into the office on Tuesday and Thursday will complete a questionnaire, or all clients who are seen between April and June. Avoid approaching only those clients who appear "nice," thereby potentially biasing the sample. It is important to track refusals so that a compliance rate can be calculated, and if possible, to obtain some basic demographic data on those who refused.

Interviews

The second major type of survey is the interview. Interviews can be conducted in-person or over the telephone. The interviews described in this section are generally not the open-ended kind described in Section 5. This section refers to a standardized interview, where the questionnaire form functions like a script (Rosenthal & Rosnow, 1984). The questions are to be read to the respondent exactly as worded. As previously described, question-order effects are a distinct possibility. Only by administering the same questionnaire to everyone can the evaluator control for this type of effect. When question order is the same, differences observed between subjects are not due to question order but are due to some other factor (such as the intervention).

In-person interviews are the most labor intensive of all survey methods, especially if the interviewer must travel to the respondent's home. But they often yield the most complete data. In-person interviews are most feasible when combined with services (such as interviews administered by the home visitor). However, it may not be practical to combine service-provision with interviewing because of time-constraints and/or lack of training for the service provider in standardized interviewing techniques. Whether in-person or on the telephone, interviewers need to be trained in how to handle non-responses and how to keep respondents on-task. They also need practice in using and accurately administering the questionnaire (especially when the questions are complicated or have multiple parts). For most evaluators' needs, telephone interviews may be the most practical. These are described next.

Telephone Interviews

Telephone interviews are an efficient method of collecting some types of data. They are equivalent to in-person interviews and are often a reasonable alternative (Bradburn, 1983). They offer several advantages (Lavrakas, 1998). The first is quality control. An interviewer administers the questionnaires so that there are fewer missing data. They are cost-effective, especially when compared to in-person interviewing. They are also best when speed is required (e.g., information needed right away in order to complete a report).

As with in-person interviews, the interviewer needs to read the questionnaire like a script. Questions and response categories need to be as simple as possible since the respondent will not be able to see the question. For example, a question with seven possible response categories would be fairly simple to answer in a self-administered format. But the evaluator may find that limiting responses to three categories works better over the phone (Lavrakas, 1998; Sudman & Bradburn, 1982).

Interviews of both types tend to have higher compliance rates than mail surveys (Fink & Kosecoff, 1985). However, the evaluator may want to follow some of the same guidelines described above to increase participation rates. A post card or letter can be sent to the respondent describing the study, and letting them know that an interviewer may be calling. The evaluator could make a pre-interview contact to find out when a good time to call would be, and to explain the purposes of the survey and that all responses will be confidential. Respondents can also be told about any incentives that they will be offered. It is important to develop a protocol for when a family will be considered a refusal (e.g., six unreturned telephone calls), and to track the number of clients/families who refuse to participate. The evaluator might also decide to conduct some interviews in person, especially among those who are impossible to interview by telephone (Lavrakas, 1998).

Who Should Be Interviewed?

One of the first steps in survey approaches is to identify a sampling frame. The sampling frame is the operationalization of the target population-a list of potential subjects to be surveyed. Should the evaluator sample from the general population of families that program services are directed towards (such as services that might be considered for a primary prevention program)? Or would it be better to sample from an already identified population-such as families being served in an existing program? A sampling frame list can be generated from directories or membership lists (e.g., involvement in community groups). Another technique unique to telephone surveys is Random Digit Dialing, where a computer generates telephone numbers. Random Digit Dialing

may not be appropriate for some populations. For instance, Random Digit Dialing may not be appropriate if it would draw in many individuals who have the same prefix but do not represent families in the program. On the other hand, it might be effective if most of the families the evaluator is interested in live within certain prefixes. (For detailed information on telephone surveys, please see Lavrakas, 1998). A technique that goes hand-in-hand with telephone interviews-computerized interviews-is described next.

Computerized Interviews

Until recently, interviewers completed questionnaires that were printed on paper, and eventually coded into computerized databases. Many large-scale telephone surveys are now conducted using Computer-Assisted Telephone Interviews (CATI). With CATI, the interviewers use a computer terminal to administer and record survey questions and responses. The survey questions appear on a computer screen, the interviewer reads the questions to the respondent, and then uses the keyboard to directly enter the respondents' replies as they are given (Lavrakas, 1998).

One advantage of CATI is increased quality of data. The CATI interviewer works from a computer screen, which is programmed to show questions in a planned order. Therefore, it is more difficult for the interviewer to accidentally omit questions or ask them out of sequence. Often, answers to some questions require using "skip patterns" (e.g., "if the answer is 'yes' skip to question 10"). The correct branching can be done automatically with CATI. In telephone interviewing without the use of CATI, following skip patterns incorrectly can be a source of error.

An additional advantage is improved monitoring of the data collection process. Because data are instantly computerized, a data set can be created and data can be analyzed early in the process. This can allow early detection of problems with the process, including problems with questions or interviewers.

Cost savings may not be a benefit for small, non-repeated surveys, due to the cost and labor involved in programming the questionnaire. However, the cost per interview with CATI decreases as sample size increases -- so in large and/or repeated surveys, CATI becomes a more cost-efficient survey method, competitive with conventional telephone methods. As this technology becomes more widely available, computerized questionnaires may become an option from which to choose.

Incentives

To thank respondents for their time, the evaluator may decide to give them a token of appreciation. This could include a certificate indicating their participation in the study. If the program involves children, the evaluator might consider offering gifts for them (e.g., small toys or books, T-shirts). Some other ideas include gift certificates for local restaurants or fast-food establishments, or movie tickets, or vouchers for other types of goods or services.

Survey Etiquette

These etiquette and ethics issues are common to all surveys. First, the interviewer should always thank the respondent for their willingness to participate. Second, out of respect for their time, the survey should be kept as short as possible and include only what is absolutely necessary. The interviewer must be sensitive to the needs of respondents at all times during the interview, and

alert for any sign that they are uncomfortable. If respondents are completing the survey in person, refreshments could be made available. Respondents may also need something to distract and entertain their children while they are answering the questions. To aid in future follow-up, it would be appropriate to ask respondents if they could be contacted again, and locator information collected, as described in Section 3 (Subsection "Retention"). At both the beginning and end of the interview, the interviewer must make certain that the respondent understands that his or her responses will be confidential. This confidentiality reminder is also helpful just before questions are asked on sensitive topics. Respondents need to be informed about how their data will be used, who will have access to it, and that it will only be reported in aggregate form (e.g., a summary of answers across respondents). At the end of the survey interviewers can close on a positive note by thanking respondents again, and giving any incentives that were promised.

In this section, information on how to develop specific, tailored survey instruments was provided. Standard measures may also be added as part of the data collection for outcome assessment.