Journal of Official Statistics, Vol. 17, No. 2, 2001, pp. 249–265

A Theory-Guided Interviewer Training Protocol Regarding Survey Participation

Robert M. Groves[1] *and Katherine A. McGonagle*[2]

A theory of survey participation suggests that sample individuals engage in more thorough cognitive processing of the survey request when their concerns about the request are addressed by the interviewer. When the concerns are satisfactorily addressed, the interview becomes a more attractive option; when they are not, a refusal tends to occur. This theory has implications for the training of interviewers in recruiting sample individuals to be respondents. A training regimen was constructed that assembled concerns perceived by senior interviewers to be common, taught trainees to classify concerns (using the terminology of the respondents) into themes, taught trainees facts to communicate regarding those concerns, and drilled the trainees in rapid, natural delivery of those facts using terminology compatible with that of the sample person. Two experimental tests of the training regimen show increases in cooperation rates for interviewers who receive the training.

Key words: Nonresponse; interviewer training.

1. Introduction

Survey designers possess two principal approaches for increasing response rates to limit the risk of nonresponse errors: features of the survey design affecting participation, and training affecting the behavior of the interviewers. The research literature regarding features of the request protocol includes studies of advanced letters (Traugott, Groves, and Lepkowski 1987; Luppes 1994), callback rules (Weeks, Kulka, and Pierson 1987), incentives (Singer, VanHoewyk, Gebler, Raghunathan, and McGonagle 1999), respondent rules (Forsman 1993), interview length (Bogen 1996), and mode of data collection. Relative to the wealth of methodological inquiry into those design features, the studies of interviewer training are quite limited (see Fowler and Mangione 1990; and Billiet and Loosveldt 1988). Further, systematic examination of interviewer training protocols regarding nonresponse shows that there has been much more attention devoted to question delivery tasks than to recruitment challenges of the interviewer (Miller-Steiger and Groves 1997).

This article describes an effort to apply a conceptual framework about survey participation to the design of an interviewer training protocol, focused on reducing nonresponse

[1] University of Michigan and Joint Program in Survey Methodology, Ann Arbor, MI, U.S.A. and College Park, MD, U.S.A.
[2] University of Michigan, ISR, Ann Arbor, MI, U.S.A.

Acknowledgements: Work supported by contract J-9-J-5-0025 with the U.S. Bureau of Labor Statistics and Cooperative Agreement 43-3AEU-7-80025 with the U.S. National Agricultural Statistics Service. We express our gratitude to various collaborators, especially Richard Clayton, Richard Rosen, Eileen O'Brien, Stephen Berry, Nancy Lyon, and Judi Clemens.

rates. It begins with a brief review of a theoretical perspective focusing on training interviewers in dealing with diverse concerns of sample individuals about the survey request. A description of two replications of an experimental training workshop introduced in two different survey organizations on different surveys follows.

2. Applying a Conceptual Framework to a Training Protocol Design

2.1. *Conceptual framework*

Following earlier work (Groves and Couper 1998), we posit that the influences on survey participation arise from the social environment, social-psychological attributes of the respondent, the survey design, and interviewer attributes. These multi-level influences shape the relatively short interactions containing the presentation of the request by the interviewer and an evaluation of the request by the respondent.

The strategies the interviewer employs to seek cooperation are not only determined by the interviewer's own ability, expectations, etc., but also by features of the survey design and by characteristics of the immediate environment and broader society. Similarly, the responses that the sample person makes to the request are affected by a variety of factors, both internal and external to the respondent, and both intrinsic and extrinsic to the survey request. Two constructs regarding interviewer behavior during the interaction with householders underlie which heuristics will dominate in the householder's decision to participate. These are labeled ''tailoring'' and ''maintaining interaction.''

Tailoring. Experienced interviewers often report that they adapt their behavior to the perceived features of the sample unit. Such interviewers engage in a continuous search for cues about the attributes of the sample household or the person who answers the door (or phone), focusing on those attributes that may suggest their approach to the participation decision.

To facilitate this, expert interviewers have access to a large repertoire of cues or phrases to describe the survey request. Which statement they use to begin the conversation is the result of observations about the housing unit, the neighborhood, and immediate reactions upon first contact with a householder. The reaction of the householder to the first statement dictates the choice of the second statement to use, and so on. With this perspective, all features of the communication are relevant – not only the words used by the interviewer, but the inflection, volume, pacing, as well as, in face-to-face surveys, the physical movements of the sample person and interviewer.

Tailoring need not necessarily occur only within a single contact. Many times contacts are very brief and give the interviewer little opportunity to respond to cues obtained from the potential respondent. Tailoring may take place over a number of contacts with that household, with the interviewer using the knowledge he/she has gained in each successive visit to that household. This implies that interviewer tailoring expertise naturally evolves with experience and trial and error. Not only have experienced interviewers acquired a wider repertoire of persuasion techniques than novices, but they are also better able to select the most appropriate approach for each situation.

Maintaining interaction. The introductory contact of the interviewer and householder is a small conversation. It begins with the self-identification of the interviewer, contains some descriptive matter about the survey request, and ends with the initiation of the questioning, a delay decision, or the denial of permission to continue. Interviewers are free to apply the ''tailoring'' over several turns in the contact conversation. How to tailor the appeal to the householder is increasingly revealed as the conversation continues. Hence, the odds of success are increased with the continuation of the conversation. The expert interviewer does *not* maximize the likelihood of obtaining a ''yes'' answer in any given contact, but minimizes the likelihood of a ''no'' answer over repeated turn taking in the contact.

We believe the techniques of tailoring and maintaining interaction should be used in combination. Maintaining interaction is the *sine qua non* of tailoring, for the longer the conversation is in progress, the more cues the interviewer will be able to obtain from the householder. However, maintaining interaction is also a compliance-promoting technique in itself, invoking the commitment principle (Cialdini 1984). That is, as the length of the interaction grows, it becomes more difficult for one actor to summarily dismiss the other.

The successful application of tailoring depends on the ability of the interviewer to evaluate the reaction of the householder to his/her call, and the effectiveness of the arguments presented. Note that the expert interviewer's initial goal is to maintain interaction (avoiding pushing for the interview) as long as the potential respondent's reaction remains neutral or noncommittal. An interviewer will continue to present different arguments until such time as the householder is clearly receptive to an interview request, or there are no more arguments to present. For inexperienced interviewers the latter may occur before the former, forcing the interviewer to initiate (prematurely in some cases) the interview request.

Is there empirical support for these theoretical assertions? We find several examples. First, the 1970's and 1980's produced a series of failed experiments in telephone survey introductions. Most attempted to find the ''best'' single script to give interviewers and force them to use it; few important differences were found (e.g., O'Neil, Cannell, and Groves 1979; Dillman, Gallegos, and Frey 1976). That is, applying a single approach to each sample household did not improve the outcome. Second, when Morton-Williams (1991) experimentally compared scripted and unscripted interviewers (both trained in ''social skills'' of dealing with respondents), the unscripted behaviors produced higher cooperation rates.

Third, Groves and Couper (1998) found that how an interviewer changes his/her behavior over successive contacts affects the likelihood of eventual cooperation.

Fourth, there is also some support from common training procedures that the ''maintaining interaction'' model operates as posited. Interviewers are typically warned against unintentionally leading the householder into a quick refusal (Miller-Steiger and Groves 1997). If the person appears rushed, preoccupied by some activity in the household (e.g., fighting among children), the interviewers are often instructed to seek another time to contact the unit. For example, a common complaint concerning inexperienced interviewers is that they create many ''soft-refusals'' (i.e., cases easily converted by an experienced interviewer) by pressing the householder into a decision prematurely.

Fifth, interviewer focus groups reveal the expert interviewers are self-aware of their techniques to tailor their comments across households (Groves and Couper 1998). Unfortunately, only rarely do interviewer recruits receive training in the multi-turn repartee inherent in maximizing the odds of a ''yes'' over all contacts. Instead, they are often trained in stock descriptors of the survey leading to the first question of the interview (Miller-Steiger and Groves 1997).

Sixth, the theoretical assertions are compatible with statistical interaction effects involving incentives. People high in community involvement show reduced incentive effects in a survey concerning a subject of public concern (Groves, Singer, and Corning, 2000); people whose past behavior demonstrates higher interest in the survey topic show reduced incentive effects (Baumgartner and Rathbun 1997). In short, influences on the decision to participate vary over individuals.

2.2. *Using the conceptual structure to design training protocols*

The conceptual structure above implies that interviewers can act to change the base propensity of a householder to participate. Altering the kind of information they provide to the respondent, emphasizing one aspect of the design or intent of the survey over others, can influence householders to participate. We believe that the most effective changes in interviewer behavior are those shaped by real concerns revealed by householders. This use of tailoring has the power to change the calculus of the decision-making on the part of the householder.

The conceptual structure also implies that interviewer training regimens might be structured to give trainees skills in tailoring. We deduce that the training might have four steps: a) the assembly of householder concerns about the request, using the native nomenclature of the population (e.g., ''I'm really busy right now!,'' ''The government has no business asking me about those things!''), b) development of alternative kinds of information relevant to those concerns (e.g., ''This will only take a few minutes of your time''), c) training of interviewers to classify householder utterances into different categories (e.g., ''Time burdens'' as a category of concerns), and d) the training of interviewers to provide, quickly and in words appropriate to an individual householder, responses to householder comments and questions (e.g., ''What would be a better time for you?'').

This suggested process has some similarity to the practice in telephone and field surveys of providing interviewers with a stock set of questions (and their answers) posed by householders. It differs, however, in two important respects. First, it provides interviewers with hundreds of different expressions of the same concerns. The training focuses on classifying diverse statements into a set of themes, each of which leads to different interviewer behaviors. Second, the process emphasizes quick interviewer response to the displayed behavior. Thus, the training must work on speed of classifying a comment into an appropriate category and on delivering an appropriately phrased response.

The training technique that seemed well-suited to these needs is repeated simulation of hypothetical householder-interviewer interactions. Initial training might concentrate on classifying behaviors into different categories of householder concerns about the survey request. Later training would grade both appropriateness of classification and speed of delivery and wording of response.

2.3. *Steps of the training regimen development*

The development and implementation of the training workshop involved three steps:

1) Focus groups of experienced interviewers were formed in order to obtain from them

examples of the actual words used by reluctant respondents to describe their concerns about the survey request. Focus group moderators sought to maximize the number of different types of concerns forwarded by the interviewers. Hundreds of utterances from respondents were collected.

2) After assembly and unduplication of the concerns, senior interviewers and training staff classified the concerns into thematic sets (e.g., ‘‘concerns about privacy,’’ ‘‘insufficient time’’), and then identified desirable behaviors of interviewers to address the concerns. There were often alternative behaviors that were judged to be of similar value by the expert interviewers in response to a specific utterance; each, however, addressed the central concern of the respondent.
3) The training workshop itself consisted of training in five skills;
    a) learning the themes of sample persons’ concerns,
    b) learning to classify sample person’s actual wording into those themes (the diagnosis step),
    c) learning desirable behaviors to address the concerns,
    d) learning to deliver to the sample person, in words compatible with their own, a set of statements relevant to their concerns, and
    e) increasing the speed of performance on b)–d). (Trainers delivered utterances exemplifying diverse themes, demanding that a trainee respond quickly, moving rapidly among the trainees at a progressively faster rate.)

The training workshop focused entirely on behaviors that might be related to the decision to participate in the survey on the part of the sample person. Thus, it was an addition to training in general interviewing techniques and specific training regarding the measurement protocol of a particular survey. In some sense, it was a ‘‘survey participation module’’ of a larger training regimen.

Our primary interest was seeing whether the training workshop on average improved interviewer cooperation rates. The experiments also offered a convenient platform to investigate a first effort to measure the degree to which the skills were actually learned by each interviewer. We had hopes that if such an evaluation tool were available, it could be used to guide decisions about whether individual interviewers were ‘‘ready’’ for production interviewing. The remainder of the article presents two experimental tests of the training protocol, both of which, for reasons of available research funding, were embedded in telephone surveys of business establishments and farm operations.

3. Experiment One

The first experiment evaluating the training approach was part of a program of research attempting to increase the rate at which employers agree to participate in a longitudinal survey. The survey, the Current Employment Statistics (CES) program of the U.S. Bureau of Labor Statistics (BLS), asks each sample employer to report six items concerning the counts and payroll for various types of employees. The request protocol involves three steps: 1) a ‘‘sample refinement’’ step where telephone interviewers contact the employer to identify an appropriate contact person and verify a mailing address, 2) an advance mailing package sent to the contact person describing the survey, and 3) an enrollment call to the contact person made to request the participation of the employer in the survey. The experiment focused only on the third step.

3.1. Design of experiment

The research design was a pre-post comparison of a set of BLS interviewers' performances. Thus, each subject (an interviewer) acted as his/her own control, in some sense, and the experimental stimulus was replicated on each interviewer. Three hundred and twenty sample employers were contacted in pre-training and 329 in post-training. Samples in both phases were stratified by four U.S. states (California, Florida, North Carolina, and Pennsylvania) and eight employer sizes, defined in terms of number of employees (0–4, 5–9, 10–19, 20–49, 50–99, 100–249 and 250+). There were no significant differences in the distribution of firm sizes between the two phases. The pre-training data collection phase lasted from February 10 until April 1, 1997. The training workshop was conducted on April 2 and April 3, 1997. The post-training data collection phase took place between April 7 and May 23, 1997.

3.1.1. Interviewers

Sixteen interviewers from one centralized telephone facility participated in both phases of the sample refinement and solicitation stages. All had been employed primarily as ongoing data collectors on the longitudinal survey but lacked experience with initial solicitation of sample firms into the CES prior to this study. Their average tenure as BLS interviewers was nearly two years, ranging from approximately one year to nearly five years. Before the pre-training phase commenced, interviewers received a general training on data solicitation that entailed a review of a sample solicitation script, multi-establishment classification rules and the respondent packet materials. This training was traditional to the facility and used techniques common to the response rate training of many organizations.

3.1.2. Interviewer-respondent interaction

At the ''sample refinement'' step, interviewers initially attempted to identify the Head of Payroll, or ''the person in charge of payroll'' to act as the contact person. All respondents were told that the contact person would receive an information packet and a follow-up telephone call. Respondents also answered a series of questions to determine their multi-establishment status, number of employees, payroll period, and industry type.

An interviewer recontacted each establishment within approximately one week of mailing the information packet. During this call, termed ''enrollment,'' the interviewer verified that the contact person had read the packet of materials and reviewed the purpose of data collection. If a new contact person was identified, the interviewer requested the name, address, title and telephone number information and attempted to reach that person. Otherwise, the interviewer collected data or made an appointment for data collection with the original contact person.

3.1.3. Interviewer training workshop

Three staff members from the University of Michigan conducted the interviewer training workshop. The workshop occurred over the course of one and one-half days. Specific topics covered in the workshop were: general principles of refusal aversion; nonspecific respondent concerns; time and burden concerns; government concerns; dealing with hostile respondents; company policy concerns; confidentiality concerns; and the ‘‘pass-off’’ (one contact wishing to transfer the interviewer to another person in the company).

The workshop interspersed mini-lectures on each topic with written and oral practice exercises, including role-playing in pairs (interviewer-respondent) and small groups. Role-playing helped the interviewer provide increasingly quick and articulate diagnoses and rebuttals. Interviewers were encouraged to practice rebuttals using their own words. Interviewer evaluations indicated enjoyment ratings of the role-playing were low, but usefulness ratings were high.

3.1.4. Evaluation of interviewers’ mastery of the workshop

Interviewers were evaluated on their understanding and mastery of the workshop materials at the close of day 1 and day 2. The goal of the day 1 evaluation was to assess the interviewer’s mastery of the diagnostic material. This exercise required interviewers to match respondent concerns with the concern type (e.g., ‘‘government concern’’ is the correct concern type match for ‘‘this is another waste of taxpayer money’’). The average score on this exercise was 13.4 out of 16 total points. The goal of the exercise on day 2 was to assess mastery of both making a diagnosis and applying a rebuttal in a time pressured setting. This exercise required interviewers to listen to a concern read aloud by one of the trainers, and record a diagnosis and rebuttal. The average score on this exercise was 11.7 out of 17 total points. The average total score (and median) on both exercises was 25 points based on 33 total points.

3.2. Refinement and enrollment results

The outcome examined is the percentage of contacted eligible sample units that cooperated with the survey request. Sample cooperation rates were significantly higher in the seven weeks following the training workshop, during Phase 2 (72.8 percent, SE = 2.9 percent, reflecting clustering by interviewers) than in the seven weeks before the workshop (62.8 percent, SE = 3.1 percent, reflecting clustering by interviewers) by a margin of approximately ten percentage points ($t = 2.32$, $p = .021$).

Figure 1 plots the interviewer-level cooperate rates pre-training and post-training. Each mark on the figure represents an interviewer’s individual cooperation rate. The fact that variation in enrollment rates declined after training (i.e., among-interviewer variance in cooperation rates falls by 73 percent (.041 to .011)) may mean that the workshop helped the lowest performing interviewers but had little effect on those interviewers who were already successful. In order to examine this possibility, the average enrollment rates post-training and the average change between the phases was compared between the group of interviewers having pre-training enrollment rates below the median with those having enrollment rates above the median. Interestingly, there was no significant difference between these two groups in post-training enrollment rates. The lower performing interviewers ‘‘changed’’ significantly more than the higher performing interviewers across the phases. Among the low performers there was an increase of nearly 24 percentage points, compared to 0 percentage points for the high performers ($t = -3.11$, $p < .007$,

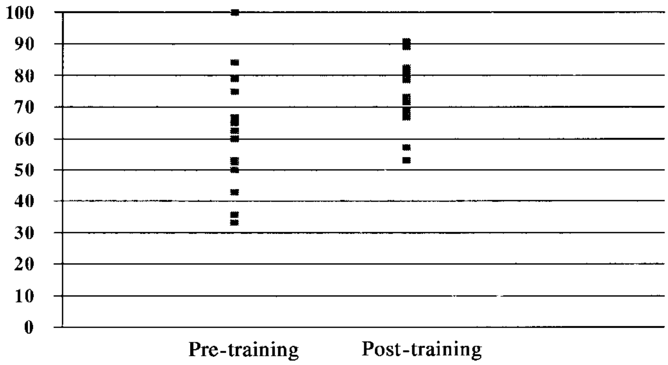

Fig. 1. Pre-training and post-training interviewer cooperation rates – experiment

reflecting interviewer clustering). The workshop, in part, allowed the lower performing interviewers to ''learn'' the skills better interviewers acquire through experience; the effect on already high performing interviewers was negligible.

3.3. Multivariate analysis of enrollment rates

While the simple pre-post comparison showed evidence of the training success, we were concerned about uncontrolled variation in the difficulty of the workload both pre-training to post-training and among interviewers. The appropriate statistical model to reflect this is a random coefficients' model which permits statistical control on other covariates of cooperation and permits the effect of the training to vary across interviewers. Finally, the structure of the model should permit tests of whether the evaluation instrument predicted which interviewers showed most gains from the training. We fit the following model,

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{1j}(POST_{ij} = 1) + \sum_{k=2}^{5}\beta_k(SIZE_k = 1) + \varepsilon_{ij}$$

where p_{ij} is the probability that the ith employer assigned to the jth interviewer cooperates with the request; $POST_{ij} = 1$ indicates that the case was sampled after the training workshop; and there are four dummy variables for the five employer size categories (0–4, 5–9, 10–19, 20–49, and 50 or more employees). These were fit as covariates, to reduce the variance of the error term and control on any unexpected differences between the pre-training and post-training samples. Note that some of the coefficients in the model have j subscripts, indicating that we expect both the intercept (the pre-training cooperation propensity) and the change from pre-training to post-training to vary across the interviewers. The effects of employers' size are posited to be fixed across interviewers.

Associated with the random coefficients are two other (interviewer-level) models: $\beta_{0j} = \gamma_{00} + u_{0j}$ and $\beta_{1j} = \gamma_{10} + u_{1j}$. That is, we specify in this base model that we have no hypotheses about why there is variation across interviewers in the pre-training performance or the training effects. In a second ''learning'' model, we specify that interviewer variation in the pre-post change is a function of how successful the interviewer was prior to training and the grade that he or she received on the training material.

Table 1. Coefficients and variance components for hierarchical logistic model predicting post-training likelihood of cooperation: Experiment One

	Base model			''Learning'' model[a]		
	Coeff	Ste	*p*	Coeff	Ste	*p*
Fixed effects (employer-level)						
0–4 employees	–			–		
5–9 employees	.48	.31	.139	.51	.31	.121
10–19 employees	–.12	.27	.669	–.11	.27	.687
20–49 employees	.50	.31	.125	.52	.31	.114
50+ employees	.88	.38	.033	.91	.38	.031
Fixed effects (interviewer-level)						
Predicting: Pre-training propensity						
Intercept (γ_{00})	–.090	.41	.831	–.090	.43	.836
Predicting: Pre-post difference in propensities						
Intercept (γ_{10})	.43	.21	.065	–.40	.46	.405
Interviewer grade on examination (γ_{11})			.070	.036	.067	
Variance components among interviewers						
Pre-post difference	.069			.11		

[a] ''Learning Model'' adds to the base model a parameter measuring the extent to which interviewers with high grades on the post-training examination achieve greater increases in cooperation after training (relative to their pre-training performance).

That is,

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(GRADE)_j + u_{1j}$$

We hypothesize that interviewers who get high grades on the training examination should have greater effects of the training (γ_{11} should be positive). We should also see that the variation in the u_{1j} across interviewers is smaller in the second model than the first.

In order to improve the ability to detect the training effects, we fit the models with the covariate, number of employees in the unit, which is influential on cooperation rates. Table 1 presents various statistics from the base model and the ''learning'' model. The base model allows us to measure the net pre-post change, controlling on the covariates. The average effect is $\gamma_{10} = 0.43$, reflecting an overall gain in cooperation propensity pre-training to post-training, falling below traditional levels of statistical significance ($p < .065$), in the presence of the size covariates. The variation among interviewers in residuals for the post-training effect is .069.

Now we move to the learning model. We first examine the coefficients for variables that attempt to explain why interviewers vary in cooperation propensities. The benefits of training are higher for those who learned the lessons of the training well (as measured by the grade on the examination, $\gamma_{11} = .070$, $p < .067$). This coefficient achieves a larger than desired p-value; we suspected that this is because of the small number of interviewers in the study. However, the coefficients displayed precisely the pattern hypothesized.

Reflecting these modifiers of the training effect through the ''learning'' model explains little of the variation across interviewers in the remaining pre-post differences. The variance component across interviewers in the training effect actually increases from .069 to .11. In short, there is little evidence that the grade on the examination successfully identified interviewers who varied in the impact of the training.

In the first experiment there was no possibility of comparing the experience of the trained interviewers with that of a comparable set of interviewers who did not undergo the training workshop. Thus, we were concerned that, although cooperation rates were improved after the training, the pre-post differences were merely the result of increased experience of all interviewers in dealing with reluctant respondents. We also sought to expose a greater number of interviewers to the training protocol because of the insufficient power of the first experiment to detect the effect of higher grades on the training examination. We achieved these goals with Experiment Two.

4. Design of Experiment Two

The second experiment was mounted within the context of the U.S. Census of Agriculture, conducted by the National Agricultural Statistics Service (NASS). This census uses a mailed, self-administered form to seek a variety of information about farm operations. Those operators who do not return a completed questionnaire are followed-up by telephone, and the census data are collected by phone if the interviewer successfully obtains cooperation. Telephone interviewers engaged in this follow-up activity were the focus of the second experiment. Thus, the survey population changed from business establishments to farm operators and from first-time solicitation to nonresponse follow-up.

Workshops were conducted in five states (Michigan, Washington, South Dakota, Georgia, and Oklahoma) between February 17 and March 24, 1998, by the same members of the University of Michigan staff who conducted the BLS training workshop. Training was conducted when approximately half of the follow-up cases were completed. The use of two random half samples ensured that equally difficult and easy cases were delivered before and after the training. Five control states, where a training did not occur, were identified (California, Wisconsin, Arkansas, North Dakota, and Alabama), and a ''pre-'' and ''post-'' time period was defined to be comparable to the preperiod and postperiod of the experimental states. In the experimental states, 96 interviewers were assigned a total of 10,559 cases combined pre-training and post-training. In the control states, 99 interviewers were assigned 12,596 cases combined pre-training and post-training the equivalent work period. Training occurred over a 2-day period, generally in the evenings for approximately 4–5 hours per session.

Interviewer Training Workshop. Each interviewer received a workshop manual with training presentation notes. Training topics included: general principles of refusal aversion; time and burden concerns; government concerns; confidentiality concerns; farm/ranch crises and changes; small operator concerns; existence of a policy against surveys; burnout from repeated survey requests; and ''pass-off'' to others. The workshop format was essentially the same as that of the first experiment, although the themes of reluctance were customized to the farm operator population faced with an Agricultural Census request.

Evaluation of Interviewers' Mastery of the Workshop. Interviewers were evaluated on their understanding and mastery of the workshop materials at the close of day 1 and day 2. The goal of the day 1 evaluation was to assess the interviewer's mastery of the diagnostic material. This exercise had interviewers match respondent concerns with the concern type, e.g., ''confidentiality concern'' is the correct concern type match for ''I'm afraid the IRS (the tax agency) will get this information.'' The average score on this exercise was 15.2 out of 16 total points. That is, there was near perfect mastery of the material from the first day, as measured by the evaluation.

The goal of the evaluation on day 2 was to assess mastery of both making a diagnosis and applying a rebuttal in a time-pressured setting. This evaluation had interviewers listen to a concern read aloud by one of the trainers, and record a diagnosis and rebuttal within one minute. The average score on this exercise was 15.3 out of 20 total points. The average total score (and median) on both exercises was 30.4 points, based on 36 total points.

4.1. *Effects of the experimental treatment*

When it comes to the results of Experiment 2, shown in Table 2, cooperation is defined as the number of completed interviews divided by the sum of complete and partial interviews and refusals. Refusals are defined using several of the codes, including ''outright'' refusals, as well as ''softer'' responses that could be considered refusals, such as ''Respondent promises to send in form.''

There appear to be larger increases (between pre-periods and post-periods) in cooperation rates in the experimental states than in the control states (16.3 percent versus 2.7 percent,

Table 2. Mean cooperation rates[a] among interviewers by state, pre-training and post-training: Experiment Two

	No. of interviewers	Pre-training		Post-training		Difference	
		Mean %	SE	Mean %	SE	Diff.	SE
Experimental states							
Michigan	12	18.6	(3.5)	62.3	(5.2)	43.7	(4.7)
Washington	13	76.2	(4.7)	69.7	(6.8)	−6.4	(5.2)
South Dakota	20	62.2	(2.4)	72.0	(2.0)	9.8	(3.1)
Georgia	24	48.6	(4.5)	74.8	(3.6)	26.2	(2.8)
Oklahoma	27	52.6	(3.8)	60.5	(3.8)	7.9	(2.6)
Total[b]	96	51.6	(1.7)	67.9	(2.1)	16.3	(1.7)
Total w/o MI[b]	84	59.9	(2.0)	69.3	(2.0)	9.4	(1.8)
Control states							
California	28	59.2	(3.5)	44.8	(3.7)	−14.4	(3.3)
Wisconsin	21	36.4	(5.4)	51.2	(4.4)	14.8	(2.9)
Arkansas	20	70.7	(4.8)	60.9	(5.9)	−9.8	(3.4)
North Dakota	18	64.6	(4.1)	64.8	(3.2)	0.2	(3.7)
Alabama	12	46.5	(6.6)	69.2	(3.2)	22.7	(4.9)
Total Control[b]	99	55.5	(2.2)	58.2	(1.9)	2.7	(1.6)

[a] Cooperation rate = No. of interviews/(No. of interviews + refusals), excluding cases receiving code of ''forms received.'' Includes only interviewers with pre-training and post-training activity.

[b] Unweighted average rate across states. Standard errors reflect clustering of observations into interviewer assignments.

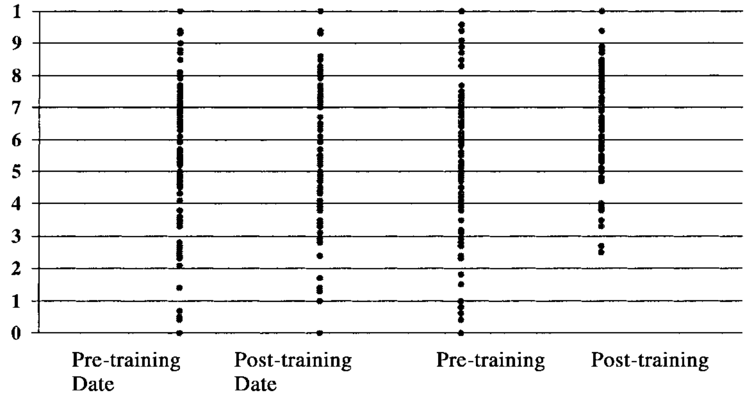

Fig. 2. Interviewer-level cooperation rates for control and experimental group interviewers by pre-training and post-training periods – Experiment Two

respectively). However, Table 2 reveals state differences in cooperation rates. Given the design, this is the combined effect of different base response propensities across state operator populations, different skills of interviewer staffs across states, and differential effectiveness of the training workshop across states. These three potential causes are confounded in the design. Michigan is an outlier state among experimental states, exhibiting very large positive effects of training (an increase in cooperation rates of 44 percentage points). When Michigan results are deleted from the comparison, the average increase in cooperation rates is about 9 percentage points in the experimental states (SE = 1.8 percentage points, reflecting interviewer clustering), compared to 2.7 percentage points in the control states (SE = 1.6 percentage points, reflecting interviewer clustering). The 9 percentage point increase in Experiment Two is to be compared with the 10 percentage point increase in Experiment One. The value of the control group is that we can estimate that about a quarter of that gain is due to the benefits of increased experience, not the training.

Figure 2 presents interviewer level cooperation rates for different groups. Here one can see that for the control interviewers, there was some reduction in among-interviewer variance in cooperation rates between the pre-experience and post-experience (a 16 percent reduction, from .053 to .044). This too reflects, we suspect, natural learning processes due to trial and error efforts of interviewers over many cases. The interviewers who experienced the workshop, however, show a much larger pre-post reduction in variation in cooperation rates (a 50 percent reduction, from .055 to .028). As before, those interviewers with the lowest pre-training cooperation rates show the largest gains.

4.2. *Multivariate analysis of Experiment Two*

The fact that larger pre-post differences are found for the experimental states than the control states was indeed the desired result of the training workshop. However, we wished to introduce statistical controls on differences in workloads among interviewers. As we did in Experiment One, we fit a model measuring the pre-post changes in performance, net of important sample differences in workload assignments. The model also permitted us to examine whether interviewers with higher grades on the valuation instrument showed larger training effects. Thus, we fit a model as a case-level model:

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{1j}(POST_{ij} = 1) + \beta_2(FORM_{ij} = 1) + \varepsilon_{ij}$$

As in Experiment One, we searched for covariates to improve our ability to estimate the training effects. The ''form'' variable reflects variation in the burden of the requested interview. Two questionnaire forms were used in the U.S. Census of Agriculture; FORM = 1 designates the shorter form of the questionnaire, which overall generated higher cooperation.

Then we tested two specifications for interviewer-level models; first, one that measures the net effect of the training session:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(TRAINING = 1)_j + u_{1j}$$

Then we added to the model a term to measure whether those obtaining higher grades on second day examination obtained greater benefits of the training:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(TRAINING = 1)_j + \gamma_{12}(GRADE)_j + u_{1j}$$

The model specification above differs from that used in Experiment One, in that it specifies that the pre-post difference is expected to be different for interviewers who received the training (experimental condition) than for those who did not (control condition). Table 3 shows values of estimated coefficients and variance components in a form similar to that used for Experiment One. The first model is a base model, used to overall net change pre-training to post-training; this shows net positive increases in likelihood of cooperation (.34, $p < .0001$) for all interviewers pooled. The second model, which measures the net training effect shows

$$\hat{\beta}_{1j} = 0.038 + 0.63(TRAINING = 1)_j$$

This shows a positive effect of the training session ($p < .0001$). The variation among interviewers in pre-post differences is reduced when the training experience is reflected in the model. About 15 percent of the among-interviewer variance in pre-post differences is associated with the training. The third model, the ''learning model,'' measures in addition an interaction term, measuring whether those obtaining higher grades experienced larger gains in cooperation. It shows a positive effect of examination grade on the pre-post difference in cooperation propensity. However, the p-value is .119, somewhat higher than anticipated. The prediction equation for the pre-post contrast in cooperation propensity of a farm operator is:

$$\hat{\beta}_{1j} = 0.038 - 0.15(TRAINING = 1)_j + 0.051(GRADE)_j$$

Care must be taken in interpreting the coefficients of the model above. For example, the −0.15 coefficient should be interpreted as the effect of the training for an interviewer who obtained a 0 score on the training examination, but the average score was about 15. Thus, for the average interviewer, controlling on pre-training performance, the average expected

Table 3. Coefficients and variance components for hierarchical logistic model predicting likelihood of cooperation: Experiment Two

	Base model[a]			Training model			''Learning'' model		
	Coeff	Ste	p	Coeff	Ste	p	Coeff	Ste	p
Fixed effects (operator-level)									
Form type (short = 1)	.29	.04	.000	.29	.04	.000	.29	.04	.000
Fixed effects (interviewer-level)									
Predicting: Pre-training propensity									
Intercept (γ_{00})	−.0085	.08	.918	−.011	.08	.893	−.011	.08	.891
Predicting: Pre-post difference in propensities									
Intercept (γ_{10})	.34	.07	.000	.038	.080	.635	.038	.080	.634
Interviewer exposed to training? (1 = Yes) (γ_{11})				.63	.10	.000	−.15	.51	.771
Interviewer Grade on Examination[b] (γ_{12})							.051	.033	.119
Variance components among interviewers									
Pre-post difference	.62			.53		$p \times^2$ diff <.001	.54		n.s.

[a] ''Base Model'' predicts same cooperation rate for all interviewers (control and experimental states), permitting pre-training and post-training differences in rates; ''training model'' predicts larger pre-post differences for the trained interviewers; ''learning model'' adds effects for trained interviewers of attained grade on the post-training examination.

[b] For control group interviewers, set to 0; for experimental groups, set to grade.

effect of the training was 0.653 in the hierarchical logit model. (In percentage terms, an interviewer using the short form would expect to move from a 57 percent to 58 percent cooperation over time without training, but from 57 percent to 72 percent after achieving a 15 grade after the training.)

Because the results in Table 2 showed large variation across states, we were concerned about how robust the findings were to state differences. We expected that most of the state variation was to be explained by differences in interviewer groups (already reflected in the two-level analysis). In the design, each experimental state was paired with a control state, so that there are five pairs of replications. There were no differences across the state pairs that would lead to a different model specification, although, as must occur, many of the coefficients became quite unstable because of the reduced sample size.

5. Summary and Discussion

The aim of these studies was to test a theoretically motivated training regimen for survey interviewers. They also gave us a chance to develop a proficiency examination attempting to predict future cooperation rates. The cooperation rates of interviewers in a pre-training phase were compared with rates in a second phase following a workshop. The two experiments, studying different populations, using different interviewer organizations, and making different survey requests, led to the same conclusion that the training served to increase cooperation rates. The second experiment, with a control group not receiving the training, supported the conclusion that the workshop had effects far beyond those from increased on-the-job experience.

The training regimen begins with input from the most proficient interviewers on staff. Not surprisingly, interviewer-level analyses in the experiments revealed that the workshop may be most cost effectively targeted to the lowest performing interviewers. As expected, low performers showed the most improvement between the phases, while high performers changed very little. These results indicate that the workshop leveled the playing field for the lower performing interviewers.

We have not yet succeeded in constructing an evaluation instrument for the training that is predictive of their future performance gains. The examination used does predict later performance, but not with the power desirable for use in deciding whether an interviewer is ready to begin production interviewing. There is an obvious (alas, now in retrospect) fix for this, which we recommend. The tests used in two experiments were timed written tests, permitting the interviewer to write down a next conversational turn in response to some oral statement by a sample person. The interviewer was given two minutes for each statement. Future examinations should involve oral presentations of simulated respondent concerns and force trainees to respond orally quickly.

Finally, the findings above have various limitations that should temper the enthusiasm of the reader. First, the theoretical developments underlying the training regimen had their genesis in face-to-face household interview surveys, where norms permit multi-turn introductory conversations between interviewers and householders. The research reported above shows potential applicability to establishment telephone surveys. We are skeptical of a naive application of these findings to RDD household telephone surveys because of the tendency for radically reduced length of interactions prior to a householder's decision. Given the success reported here, however, we are more confident that face-to-face interview surveys can profit from the training technique.

The studies were embedded in ongoing surveys, permitting the identification of interviewers who were achieving high cooperation rates and were then chosen to supply the inputs to the training protocol. This, in our belief, is key to the result of increasing overall cooperation rates, by teaching the lower-performing interviewers the skills of the higher-performing ones. There is no magic bullet in this regimen; if no interviewer mentions effective behaviors in reaction to a specific respondent concern, the training protocol will not incorporate them. We note that this places a larger burden on one-time surveys, which cannot design the protocol based on past experience. (We are now attempting to use pretests and pilot studies to acquire the data base of respondent concerns in such studies.)

We further note that there may exist a core set of respondent concerns that are common to many household survey designs. We have found, however, that the initial focus groups discover unique concerns of the population studied, given the recruitment procedures chosen, given a specific topic, and given a survey with specific burden. In short, the training for recruitment must be designed for each survey separately, as must the training design for administration of the questionnaire.

We find these results encouraging, implying that the general theoretical structure reviewed above has merit. Further, this one application of the theory does not exhaust its implications for interviewer behaviors that may produce higher participation rates. As noted earlier (Groves and Couper 1998), if the survey designer can arm interviewers with more knowledge about the sample case prior to their contacting the case, then interviewers can construct working hypotheses about what concerns about the survey request different respondents might harbor. This argues for the use of rich sampling frames whenever possible (even alerting the interviewer to the potential privacy concerns of unlisted households in telephone surveys might be useful). Further, it argues that instructing the interviewers to record process data such as observations about whether the respondent asked questions, expressed negative reactions, or reported limited time availability might alert the behavior of the next interviewer's call to the household.

6. References

Baumgartner, R. and Rathbun, P. (1997). Prepaid Monetary Incentives and Mail Survey Response Rates. Paper presented at the annual conference of the American Association of Public Opinion Research, Norfolk, VA, May.

Billiet, J. and Loosveldt, G. (1988). Interviewing Training and the Quality of Responses. Public Opinion Quarterly, 52, 190–211.

Bogen, K. (1996). The Effect of Questionnaire Length on Response Rates – A Review of the Literature. Paper presented at the annual conference of the American Association for Public Opinion Research, Salt Lake City, UT.

Cialdini, R.B. (1984). Influence: The New Psychology of Modern Persuasion. New York: Quill.

Dillman, D., Gallegos, J., and Frey, J. (1976). Reducing Refusal Rates for Telephone Interviews. Public Opinion Quarterly, 40, 66–78.

Forsman, G. (1993). Sampling Individuals Within Households in Telephone Surveys. Paper presented at the annual conference of the American Association for Public Opinion Research, St. Charles, IL.

Fowler, F.J. and Mangione, T. W. (1990). Standardized Survey Interviewing. Newbury Park: Sage Publications.

Groves, R.M. and Couper, M.P. (1998). Nonresponse in Household Interview Surveys. New York: John Wiley.

Groves, R.M., Singer, E., and Corning, A.C. (2000). Leverage-Salience Theory of Survey Participation: Description and an Illustration. Public Opinion Quarterly, 299–308.

Luppes, M. (1994). Interpretation and Evaluation of Advance Letters. Paper presented at the Fifth International Workshop on Household Survey Nonresponse, Ottawa.

Miller-Steiger, D. and Groves, R. (1997). Interviewer Training Techniques: Current Practice in Survey Organizations. Ann Arbor, MI: Survey Research Center.

Morton-Williams, J. (1991). Obtaining Co-operation in Surveys – The Development of a Social Skills Approach to Interviewer Training in Introducing Surveys. Working Paper No. 3, London: Joint Centre for Survey Methods.

O'Neil, M., Groves, R., and Cannell, C. (1979). Telephone Interview Introductions and Refusal Rates: Experiments in Increasing Respondent Cooperation. Proceedings of the Section on Survey Research Methods, American Statistical Association, 252–255.

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T., and McGonagle, K. (1999). The Effects of Incentives on Response Rates in Interviewer-Mediated Surveys. Journal of Official Statistics, 15, 217–230.

Traugott, M. W., Groves, R. M., and Lepkowski, J. M. (1987). Using Dual Frame Designs to Reduce Nonresponse in Telephone Surveys. Public Opinion Quarterly, 51, 48-57.

Weeks, M. F., Kulka, R.A., and Pierson, S.A. (1987). Optimal Call Scheduling for a Telephone Survey. Public Opinion Quarterly, 51, 540–549.

Received September 1999
Revised December 2000