RESEARCH REPORT SERIES
(Statistics #2013-07)

Effects of Interviewer Refresher Training and Performance Monitoring in the 2011 National Crime Victimization Survey

Joseph L. Schafer

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: October 23, 2013

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Effects of Interviewer Refresher Training and Performance Monitoring in the 2011 National Crime Victimization Survey

Research Report Series (Statistics)

Joseph L. Schafer

October 2013

RRS2013-07

Disclaimer. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

BACKGROUND

The National Crime Victimization Survey (NCVS), sponsored by the Bureau of Justice Statistics (BJS), is a principal source of information about criminal victimization in the United States. Based on a nationally representative sample of households, NCVS tracks crimes against persons age 12 and older regardless of whether the incidents were reported to police. Products from NCVS include estimated annual counts and rates of victimizations within major categories of crime, which include nonfatal violent crimes against persons (rape, sexual assault, robbery, simple and aggravated assault) and property crimes against households (household burglary, motor vehicle theft, and theft).

The U.S. Census Bureau serves as the primary data collector for the NCVS, conducting interviews and processing sample data on a monthly basis. Estimated counts and rates are released each calendar year, and microdata files are made publicly available through the National Archive of Criminal Justice Data at the University of Michigan.

The NCVS is a rotating panel survey; sampled households are interviewed every six months for a total of seven interviews over three years. Census Bureau Field Representatives (FRs) conduct interviews using a computer-assisted personal interviewing (CAPI) instrument. The initial interview for a household is always by personal visit. FRs are encouraged to conduct follow-up and subsequent interviews by telephone to reduce data-collection costs. The six-month window prior to the interview

HIGHLIGHTS

To promote high levels of performance by NCVS interviewers, an intervention of refresher training and performance monitoring was phased in by a randomized experiment. Interviewer teams were randomly assigned to two cohorts. Those in the first cohort were trained in late 2011, and after training, supervisors monitored interviewer performance using an expanded set of data quality indicators. For the second cohort, training and monitoring were delayed until 2012. Effects of the intervention were estimated by comparing outcomes for the two cohorts in the latter months of 2011, using statistical models that account for the experimental design. We estimate that the intervention

- raised the apparent rate of household property crime by 28%, a result that is statistically significant, and
- raised the apparent rate of violent crime by 19%, a result that is not statistically significant.

The effects of the intervention were larger and more significant for crimes that had not been reported to police. These effects, which were evident in the first cohort in the latter months of 2011, did not affect any of the NCVS crime estimates published for 2011, because those estimates were based only on pre-intervention interviews.

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

serves as the reference period for reporting crime incidents, and results are aggregated and weighted to produce estimates for each collection year.

During the interview, victimizations are identified and enumerated by a two-step process. The first step is a screener interview in which respondents are asked a series of questions about their experiences with crime during the previous six months. Each screener question consists of a main question "stem" followed by multiple "cues" to prompt recollection of the types of incidents relevant to the NCVS. For example, one of the stems is, "Since (start date of reference period), were you attacked or threatened OR did you have something stolen from you…" This stem is followed by the cues, "At work or school," "In places such as a storage shed or laundry room, a shopping mall, restaurant, bank, or airport," and "While riding in any vehicle." FRs are supposed to carefully read aloud all of the screener stems and cues, because they have been designed, tested and refined to jog respondents' memories to help them recall events that they might otherwise have forgotten. Failure to read the stems and cues in their entirety may cause some crimes to go unreported [1].

For each incident discovered during the screener process, the interview proceeds to the second step, a detailed crime incident report which records information about the date and characteristics of the event. Respondents are not directly asked what type of crime they experienced (robbery, burglary, etc.) because many will be unfamiliar with the legal definitions of these crimes, and these definitions may vary by jurisdiction. Rather, the incident report collects details about the event which are used later to classify it into a standardized taxonomy. For example, respondents are not asked whether they were robbed. Rather, they are asked whether they were physically present during the incident, whether they were threatened or attacked, whether a weapon was used, and if the offender took or attempted to take cash or other property from the victim. Depending on the answers, the event may be classified as a completed or attempted robbery.

A Renewed Focus on Data Quality

To maintain high levels of performance, interviewers on many surveys are given refresher training at regular intervals. Prior research has shown that training can significantly reduce interviewer error, increase adherence to best practices and nonbiasing interpersonal behaviors, reduce rates of item nonresponse, and improve measures of data quality [2] [3] [4]. Recent experience with another survey administered by the Census Bureau has shown that interviewers found training to be very helpful for reacquainting them with proper procedures, addressing issues encountered in the field, and allowing them to share their experiences and learn from their peers [5].

Flat program funding in recent years led to the elimination of refresher training for NCVS field staff. Prior to 2011, the last classroom training for NCVS interviewers took place in 2006, before the introduction of the CAPI instrument that July. Budgetary constraints also led BJS and the Census Bureau to discontinue general performance reviews for interviewers and to reduce the size of a reinterview sample which has been used to measure, monitor and control data quality.

One of the benefits of a CAPI environment is the ability to capture keystrokes and timing information for every case touched by an interviewer. Analysis of CAPI data showed that the average NCVS screener interview lasted less than two minutes, and in some cases the screener took less than a minute. These results raised major concerns about data quality, because the screener interview should at a minimum last three and a half to four minutes. Short screener times suggest that FRs are skipping cues, leading to possible underreporting of crimes.

In response to these and other concerns, the Census Bureau and BJS designed a comprehensive data quality improvement program which began in 2011. This intervention included refresher training for FRs. It also introduced new methods for assessing interviewer performance. The Census Bureau has recently added new systems for for collecting and summarizing case-level data quality variables. The contact history instrument (CHI) and perfomance and data analysis system (PANDA) provide additional data quality indicators by which supervisors can monitor staff performance. Refresher training provided the opportunity to introduce these new performance metrics.

If this intervention had been given to all FRs at the same time, the effects of the intervention would have been difficult to measure. For this reason, the program was phased in using a randomized experiment. Teams of FRs were randomly assigned to two groups. One group (Cohort 1) received the intervention during the third quarter of 2011, and the other group (Cohort 2) received it during the first quarter of 2012. Under this experimental design, the impact of the program can be measured by comparing outcomes for the two groups during the latter months of 2011 when Cohort 1 had received the but Cohort 2 had not.

Preview of Findings

Comparisons between the cohorts in late 2011 show that the intervention did impact survey outcomes. Using a statistical model that adjusts for the experimental design and random baseline differences between the cohorts, we estimate that FRs in Cohort 1 reported household property crimes at a rate that was 28% higher than the FRs in Cohort 2. The result is statistically significant. For violent crimes, the estimated increase is 19% and not statistically significant. These results do not indicate that the actual rates of crime were larger for Cohort 1. Rather, the intervention led to an increase in *reporting* of crime in Cohort 1. One plausible explanation is that the intervention, through changes in interviewer-respondent interactions, led to cognitive recall of events that would otherwise have forgotten.

We also found that intervention effects were strong and highly significant for crimes that had not been reported to police, but small and insignificant for crimes that had been reported to police. If crimes reported to police tend to be more serious and more salient in victims' memories, then this finding is consistent with previous work that has demonstrated a relationship between salience and difficulty of recall [1] [6] [7].

Partly based on these results, a decision was made to omit post-intervention Cohort 1 interviews from the published NCVS figures for 2011 [8]. Cases from post-training interviews in Cohort 1, which comprised about one-eighth of all interviews during the year, were removed from the sample prior to weighting and estimation. The rationale for removing those cases was not that the post-intervention interviews are considered unreliable. On the contrary, training and monitoring have presumably improved data quality. Rather, the post-intervention interviews were removed to keep estimates from 2011 comparable to those from previous years when no refresher training was done, averting a possible break in series. Because this action was taken, *no intervention effects are present in any of the published results from 2011.*

In this present report, we limit our attention to the effects of refresher training and performance monitoring in 2011. Because the cases affected by this intervention were ultimately removed from the 2011 NCVS, the implications of these findings are largely academic. In a companion report by Schafer (2013), however, we analyze the effects of refresher training and performance monitoring in 2012, along with other phenemena that occured since 2008 [9]. Evidence for intervention effects in 2012 cannot come from the randomized experiment, but must be gleaned from other sources. In that document, we devise a flexible class of longitudinal models that incorporates long-term trends, annual periodic trends, effects for interventions and additional covariates.

Scope of This Report

In the remaining sections of this present report, we describe the intervention of refresher training and performance monitoring and give details of the experimental design that was used to phase in the program. We devise a class of statistical models for estimating the effects of the intervention in late 2011 based on the experimental design. We analyze the effects on major categories of personal and property crime, classified by whether or not the crimes were reported to police, and on screener times and response rates. Because many of these crime categories have small base rates, the estimated training effects tend to be noisy. To reduce noise and strengthen the estimates, we pooled evidence across categories of crime by a special meta-analysis procedure that takes into account the overlapping nature of the categories. We discuss the implications of our findings for interpreting estimates from the 2011 NCVS. Details of our computational methods are provided in a technical appendix.

IMPLEMENTATION OF THE DATA QUALITY INTERVENTION

Experimental Design

Because FRs were assigned to cohorts randomly, the phase-in period for the intervention may be regarded as a randomized experiment with two treatment regimens. A review of randomized experiments within surveys is provided by van den Brakel (2008) [10].

Each FR is employed at a Census Bureau Regional Office and works on a team managed by a Senior Field Representative (SFR). The training that began in 2011 included management strategies to encourage and maintain adherence to interview procedures. Thus the experimental units to which the treatments (intervention in 2011 versus intervention in 2012) were applied are SFR-led interviewer teams, not individual FRs. To help ensure that the two cohorts would be well balanced with respect to geography and other important characteristics, the teams within each Regional Office were grouped into homogeneous pairs based on prior rates of victimization, screener interview times, and monthly workload. Within each pair, one team was randomly selected and assigned to Cohort 1, and the other team was assigned to Cohort 2. This experimental design is an example of what health researchers call a matched-pair cluster-randomized trial [11] [12].

The use of matched pairs in cluster-randomized trials has a mixed reputation. Some have argued that matching is counterproductive and that methods for analyzing matched designs are problematic [13]. More recently, however, it has been shown that those warnings are unfounded; matching can substantially increase efficiency, especially when the cluster sizes vary [14].

The matching and randomization procedure created two cohorts that had similar characteristics when the experiment began. Without randomization, there would have been no way of knowing whether a difference in outcomes between the cohorts was a bonafide effect of the intervention, or merely an artifact of pre-existing systematic differences between the cohorts on variables related to the outcome. The models that we describe and fit in the sections ahead provide formal tests that demonstrate that the two cohorts were well balanced, with no significant differences at the start of the experiment, on any of the outcome variables we considered.

When analyzing data from a clustered trial, it is essential to take the clustering into account; failure to do so may distort treatment effects and significance levels [15]. Hill and Scott (2009) analyzed matched-pair cluster-randomized trials using a family of hierarchical linear models (also known as multilevel regression or mixed-effect models) with a fixed coefficient for the treatment effect and varying coefficients for the cluster pairs [16]. We have adopted an approach similar to theirs, but with two important modifications. First, we expanded their family of linear regressions to include loglinear and logistic models, which are more appropriate when the outcomes are rates and proportions. Second, our models are explicitly longitudinal; they control for random pre-treatment differences between the cohorts, which increases the precision of between-cohort comparisons.

Training Procedures

In July 2011, The Census Bureau and BJS held a conference with field survey supervisors to discuss survey administration topics and the NCVS data quality improvement program. On the last day of the conference, Census conducted a "train-the-trainer" session, which was a dry run of the verbatim refresher training materials. The theme of the refresher training was "a renewed focus on data quality." The training included:

- a reintroduction to the NCVS;
- a review of the screener questions, crime incident report, and contact history instrument, and concepts and definitions;
- paired practice interviews;
- topics specific to Regional Offices;
- the introduction of twelve new data quality indicators (DQIs); and
- a review of the pre-training knowledge test.

Following the train-the-trainer session, survey supervisors conducted multiple training sessions in local areas within their respective regions. These sessions lasted a day and a half. The vast majority of FRs in Cohort 1 were trained in August (89%); the rest were trained in July (3%), September (6%), and October (1%). FRs from Cohort 2 were trained in January (2%), February (92%), March (3%) and May (3%) of 2012. In December 2011, prior to the Cohort 2 trainings, Census conducted a video conference and additional train-the-trainer session with Regional Office staff.

Performance Monitoring

Starting with the September 2011 interviewing, field survey supervisors began providing monthly feedback to Cohort 1 interviewers based on the set of twelve DQIs presented in refresher training. Reports were generated in the PANDA system to track the performance of interviewers related to each DQI. Cohort 2 interviewers began to appear on the reports in April 2012. Until FRs were trained, supervisors were instructed to manage them in the same manner that all staff were managed before refresher training began. After FRs were trained, supervisors used the DQI reports to evaluate interviewer performance, identifying areas that required improvement and providing feedback designed to meet survey quality standards. The DQIs are measures of

- household and person-within-household response rates,
- completeness of the screener questions and crime incident report data items,
- whether any crime incident report items had to be changed during the editing/coding process,
- completion of the contact history instrument record,
- screener times and crime incident report times,
- interviews that took place outside the monthly data-collection period, or cases where no contact was made until after the 15th of the interview month, and
- interviews that began between 10 pm and 7 am.

A Limitation of this Design

One feature of this experimental design is that the two major components of this intervention, refresher training and performance monitoring, were enacted for each cohort at roughly the same time. Consequently, the effects of these components are confounded. There is no way to estimate the effect of performance monitoring in the absence of refresher training or vice versa. This confounding has important ramifications for designing future data quality interventions. As performance monitoring continues in years ahead, it is unclear what the results from this experiment may portend for additional refresher training.

MODELING THE EFFECTS OF THE INTERVENTION ON THE REPORTING OF CRIME

Cross-Product Ratios

To measure the effects of the intervention, we divided the 2011 calendar year into pre- and post-training periods, which we call Time 1 and Time 2, respectively. Most interviewer teams in Cohort 1 received their training in August, but some were trained later. For each FR in Cohort 1, we defined Time 1 as the calendar months up to and including the month in which the team's training began, and Time 2 as the months after training. Although FRs in Cohort 2 were not trained in 2011, we divided their year in a similar fashion, defining Time 1 as January through August and Time 2 as September through December.

Table 1 shows the number of crimes against persons recorded by FRs in 2011 by cohort and time period, along with the number of persons interviewed. Dividing the number of crimes by the number of interviews gives a raw crime rate. The rates in Table 1 are substantially smaller than the annual victimization rates for personal crime that appear in NCVS reports (see [17], for example) because published estimates are weighted to account for the sample selection procedures, nonresponse, and differing windows of time. Annual rates refer to a whole year, whereas the interview refers to a period of six months. Moreover, crimes are not precisely the same as victimizations, because some crimes have multiple victims. Nevertheless, by examining the rates in Table 1, we gain some understanding of how the intervention affected the reporting of personal crimes.

Table 1: Crimes against persons in the 2011 NCVS refresher training experiment by cohort and time period

	Cohort 1		Cohort 2	
	Time 1	Time 2	Time 1	Time 2
Crimes	516	282	384	209
Interviews	60,972	28,751	52,123	26,530
Rate*	8.46	9.81	7.37	7.88

**crimes per 1,000 interviews*

FRs in Cohort 1 discovered more crimes than FRs in Cohort 2 during the same period. Dividing the rate for Cohort 1, Time 2 by the rate for Cohort 2, Time 2, we obtain

$$\frac{9.81}{7.88} = 1.24, \tag{1}$$

which suggests that the intervention increased the rate of personal crime. Some of this increase, however, can be attributed to the fact that the FRs in Cohort 1 had been reporting more crimes than the FRs in Cohort 2 when the experiment began; during Time 1, Cohort 1's rate was 15% higher than Cohort 2's,

$$\frac{8.46}{7.37} = 1.15.$$

This discrepancy between the rates at Time 1 is not statistically significant, as we will demonstrate later. The random procedure used to assign teams to cohorts ensures that, if the experiment were repeated many times, the two cohorts would on average be identical in every respect. But in this particular experiment, it happened by chance that the teams assigned to Cohort 1 had been reporting

15% more personal crimes than Cohort 2. A more precise estimate of the intervention effect that adjusts for this random pre-treatment discrepancy is the cross-product ratio,

$$\frac{9.81\,/\,8.46}{7.88\,/\,7.37} = 1.085. \tag{2}$$

After the intervention, Cohort 1's rate went up from 8.46 to 9.81. But without any intervention, Cohort 2's rate went up from 7.37 to 7.88. The cross-product ratio (2) suggests that the intervention led to an increase from Time 1 to Time 2 that was 8.5% larger than the natural increase that would have been seen without it.

A similar strategy can be applied to property crimes, which are crimes against households. Numbers of property crimes by cohort and time period are reported in Table 2, along with numbers of households interviewed and the resulting rates.

Table 2: Property crimes in the 2011 NCVS refresher training experiment by cohort and time period

	Cohort 1		Cohort 2	
	Time 1	Time 2	Time 1	Time 2
Crimes	2,103	1,215	1,842	848
Interviews	39,339	19,252	32,945	16,815
Rate*	53.5	63.1	55.9	50.4

** crimes per 1,000 interviews*

Without intervention, the rate in Cohort 2 dropped from 55.9 at Time 1 to 50.4 at Time 2. With intervention, the rate in Cohort 1 rose from 53.5 at Time 1 to 63.1 at Time 2. The cross-product ratio is

$$\frac{63.1\,/\,53.5}{50.4\,/\,55.9} = 1.309, \tag{3}$$

which suggests that the intervention increased the reporting of property crime by about 31%.

To judge whether these effects are statistically significant, we need a model that accounts for the longitudinal nature of these data, because the same FRs are contributing responses at Time 1 and Time 2. The model should also be congruent with the matched-pair cluster experimental design. To develop an appropriate model, we first recast the cross-product ratios in (2) and (3) as parameters in a loglinear model.

A Loglinear Model for Rates

Let Y_{jt} and N_{jt} denote the number of crimes and the number of interviews, respectively, in Cohort j during Time t. For example, using the data shown in Table 1, we have $Y_{11} = 516$, $N_{11} = 60{,}972$, $Y_{12} = 282$, $N_{12} = 28{,}751$, and so on. Suppose that Y_{jt} is distributed as

$$\begin{aligned} Y_{jt} &\sim \text{Poisson}(\mu_{jt}), \\ \log \mu_{jt} &= \log N_{jt} + \boldsymbol{x}_{jt}^T \beta, \end{aligned} \tag{4}$$

where μ_{jt} denotes the expected value of Y_{jt}, log is the natural logarithm, the superscript T is the vector transpose, $\boldsymbol{x}_{jt}$ is a vector of known predictor variables describing Cohort j at Time t, and β is a vector of unknown coefficients to be estimated.

Model (4) is called a loglinear model or a Poisson regression with a log link. It is an example of a generalized linear model, a widely used family that includes normal linear regression and logistic regression [18] [19]. The term $\log N_{jt}$ on the right-hand side is called an offset; it plays the role of a predictor variable with a coefficient assumed to be one. Under this model, the mean of Y_{jt} is proportional to N_{jt}, and

$$\exp(\mathbf{x}_{jt}^T \beta) = \mu_{jt}/N_{jt}$$

is the expected crime rate. The elements of β, when exponentiated, are the multiplicative effects of the predictors on the expected rate.

Suppose we define a dummy indicator for Time 2,

$$T_{2t} = 1 \text{ if } t = 2, \text{ and 0 otherwise,}$$

and a dummy indicator for Cohort 1,

$$C_{1j} = 1 \text{ if } j = 1, \text{ and 0 otherwise.}$$

And suppose we define $\boldsymbol{x}_{jt}$ as the vector

$$\boldsymbol{x}_{jt} = (1,\, T_{2t},\, C_{1j},\, T_{2t} \times C_{1j})^T. \tag{5}$$

Applying this model to the property-crime counts from Table 2, we obtained the results shown in Table 3. We fit this model using the glm command in R Version 2.12 [20], but equivalent results would be obtained from any program that fits generalized linear models by maximum likelihood. Table 3 shows the estimated coefficients, standard errors, and p-values for testing whether the coefficients are significantly different from zero.

Table 3: Coefficients, standard errors, and p-values from loglinear model for property crime

	Coef	SE	p^*
Constant	-2.884	.0233	—
Time	-0.103	.0415	.013
Cohort	-0.045	.0319	.160
Time×Cohort	0.269	.0550	.000

**based on a two-tailed normal approximation*

In this model, the effect of refresher training is measured by the Time × Cohort interaction. Exponentiating that coefficient, we obtain

$$\exp(0.269) = 1.309,$$

which, except for rounding error in subsequent digits, is identical to the cross-product ratio (3) based on the raw rates. The imbalance in crime rates between the two cohorts when the experiment began is measured by the main effect of Cohort. Exponentiating that coefficient, we obtain

$$\exp(-0.045) = 0.956,$$

which, except for rounding error, is identical to the ratio of raw rates for the two cohorts at Time 1,

$$\frac{53.5}{55.9} = 0.957.$$

Unfortunately, the standard errors and significance levels shown in Table 3 are not reliable, because the estimation procedure incorrectly supposes that the event counts Y_{11}, Y_{12}, Y_{21} and Y_{22} are independent. Independence is violated because these data are longitudinal; except for minor discrepancies due to hiring of new FRs and attrition, the same FRs are present at Times 1 and 2. Moreover, the estimation procedure fails to account for the fact that FRs are clustered within teams, and teams were assigned to cohorts using team pairs as an experimental blocking factor. These features of the study design induce correlations among the Y_{jt}'s and suggest that the standard errors shown in Table 3 are too small.

Accounting for the Study Design

To account for the study design, we expanded the loglinear model (4) to describe counts at the interviewer level. Let Y_{ijkt} and N_{ijkt} denote the number of crimes and number of interviews, respectively, for interviewer i in Cohort j and team pair k during Time t. We suppose that

$$\begin{aligned} Y_{ijkt} &\sim \text{Poisson}(\mu_{ijkt}), \\ \log \mu_{ijkt} &= \log N_{ijkt} + \boldsymbol{x}_{ijkt}^T\boldsymbol{\beta} + \alpha_i + \boldsymbol{x}_{ijkt}^T\gamma_k. \end{aligned} \quad (6)$$

In this expanded model, α_i represents an effect due to interviewer i. We assume that the interviewer effect is normally distributed with mean zero and variance σ_α^2,

$$\alpha_i \sim N(0, \sigma_\alpha^2).$$

The vector of predictors $\boldsymbol{x}_{ijkt}$ has the same form as before (5); it includes a constant, a dummy indicator for Time = 2, a dummy indicator for Cohort = 1, and a Time × Cohort product. However, the coefficients for these terms are now allowed to vary across team pairs, and we assume that the team-pair effects are jointly normally distributed,

$$\gamma_k \sim N(0, \boldsymbol{\Sigma}_\gamma),$$

where $\boldsymbol{\Sigma}_\gamma$ is a 4 × 4 matrix of variances and covariances.

A similar family of models for matched-pair cluster randomized trials was proposed by Hill and Scott (2009) [16]. Their models assumed a normally distributed response variable and linear effects for predictors. Our model, which assumes a Poisson response and loglinear effects, is more appropriate for describing rates.

The key parameters of this expanded model are contained in β; they include the Time × Cohort interaction, which measures the overall effect of refresher training, and the main effect for Cohort, which measures the degree of imbalance in crime rates between the two cohorts when the experiment began. But the model now has additional parameters σ_α^2 and $\boldsymbol{\Sigma}_\gamma$. These are not of primary interest, but they help to describe the correlations among the Y_{ijkt}'s, so that standard errors and significance levels for the key parameters become more realistic.

Our expanded model is an example of a generalized linear mixed model (GLMM) [21] [22]. Software for fitting GLMMs is available (for example, [23] [24]), but most programs have difficulty with this model, because it has a non-normal response and three levels of units (observations nested within interviewers, and interviewers nested within team pairs). These data are sparse, because crimes are relatively rare; most of the Y_{ijkt}'s are zero. Sparseness causes maximum-likelihood procedures to behave poorly unless the samples are extremely large. In order to fit this expanded model, we implemented a custom algorithm that uses Bayesian estimation and Markov chain Monte Carlo [25]. Details of the procedure are provided in the technical appendix.

ESTIMATED INTERVENTION EFFECTS

Property Crime

Applying the expanded model to property crime, we obtained the estimates shown in Table 4.

Table 4: Coefficients, standard errors, and p-values from expanded model for property crime

	Coef	SE	p^*
Constant	−3.159	.0650	—
Time	−0.158	.0710	.026
Cohort	−0.005	.0875	.926
Time×Cohort	0.291	.0875	.001

* *equal-tailed Bayesian p-value*

Comparing these results to those in Table 3, we see that the coefficient for Time × Cohort is slightly larger. The exponentiated coefficient is

$$\exp(0.291) = 1.338,$$

which suggests that intervention increased the reporting of property crime by about 34%. The standard error is wider (.0875 versus .0550); accounting for the study design has increased the uncertainty, but the effect is still significant. The main effect of Cohort is small and insignificant, demonstrating that the two cohorts were well balanced with respect to property crime when the experiment began. The estimated variance due to interviewers is

$$\hat{\sigma}^2_\alpha = 0.363,$$

and the estimated variances and covariances due to team pairs are

$$\hat{\boldsymbol{\Sigma}}_\gamma = \begin{bmatrix} .207 & -.011 & -.067 & .006 \\ -.011 & .209 & .003 & -.108 \\ -.067 & .003 & .244 & -.053 \\ .006 & -.108 & -.053 & .286 \end{bmatrix}.$$

We repeated the analysis for major categories of property crime, which include household burglary, motor vehicle theft and other theft. The estimated coefficients for Time × Cohort from these models, along with their standard errors and p-values, are shown in Table 5.

Table 5: Estimated intervention effects, standard errors and p-values for categories of property crime

	Coef	SE	p^*
Property crime	0.291	.0875	.001
Household burglary	0.359	.1774	.048
Motor vehicle theft	−0.649	.4647	.153
Theft	0.298	.0988	.002

* *equal-tailed Bayesian p-value*

For household burglary and theft, the effects are significant and positive; training increased the reporting of crime in those categories. For motor vehicle theft, the coefficient is negative and insignificant. That estimate is imprecise, with a standard error much larger than the others, because the number of observed cases of motor vehicle theft is small.

Police Reporting

Only 38% of the property crimes discovered by FRs were reported to police. Based on comments from the NCVS Data Review Panel, we hypothesized that the intervention might lead to greater recall of crimes that had not been reported to police. Fitting separate models to reported and unreported crimes, we found that the effects of training are moderated by police-reporting status. Results from these models are shown in Table 6. Except for motor vehicle theft, for which the estimates are very imprecise, a pattern has emerged. Among crimes not reported to police, the effects of the intervention are large and statistically significant, but among crimes reported to police, the effects are small and insignificant.

An intuitive explanation for this pattern is that a crime reported to police tends to be more salient in the respondent's memory and is more easily discovered by an FR during a screener interview, whereas an unreported crime is less salient and will be discovered only if the interview is done more carefully. The known relationship between saliency of the crime and difficulty of recall was a key consideration in the present design of the NCVS, including the wording of screener interviewer and the six-month reference period [1] [6] [7].

Table 6: Estimated intervention effects, standard errors and p-values for categories of property crime by whether the crime was reported to police

	All crimes			Reported to police			Not reported to police		
	Coef	SE	p^*	Coef	SE	p^*	Coef	SE	p^*
Property crime	.291	.088	.001	.130	.125	.283	.384	.104	.000
Household burglary	.359	.177	.048	.260	.228	.249	.674	.268	.010
Motor vehicle theft	−.649	.465	.153	−.704	.504	.151	−.836	1.33	.347
Theft	.298	.099	.002	.094	.152	.524	.383	.113	.002

* *equal-tailed Bayesian p-value*

Crimes Against Persons

The results presented thus far are for crimes committed against households. For crimes against persons, the effects of the intervention are more difficult to measure, because in some of these categories the numbers of crimes reported to NCVS interviewers were very small. Nevertheless, we fit models of the same form (6) for major categories of personal crime. The results are displayed in Table 7. These estimates are noisy; only one of the effects is statistically significant at the .05 level. However, for crimes not reported to police, the estimated coeffients are all positive, which is unlikely to happen merely by chance if training effects do not exist.

POOLING THE ESTIMATES

Meta-Analysis

For many categories of crime, intervention effects are poorly estimated because data are sparse. If we believe that the intervention produces similar results across categories, we may strengthen the estimates by pooling evidence across categories. The practice of combining evidence is called meta-analysis [26]. Meta-analysis is typically used to synthesize published results describing different studies of the same phenomenon. The results from each study are summarized by a few simple measures and then combined to yield an overall estimate of the common effect.

In the random-effects approach to meta-analysis [27], we suppose that the results from study i ($i = 1, \dots, N$) are summarized by a point estimate $\hat{\theta}_i$ and a standard error $\hat{\sigma}_i$. We regard $\hat{\theta}_i$ as an estimate for the true effect in study i, which is denoted by θ_i, and we suppose that $\hat{\theta}_i$ is normally distributed around that true effect,

$$\hat{\theta}_i \sim N(\theta_i, \hat{\sigma}_i^2).$$

We then suppose that the true effects are randomly distributed around an overall effect,

$$\theta_i \sim N(\mu, \tau),$$

where μ is the overall effect and τ is the between-study variance. If μ and τ were known, a strengthened estimate for θ_i would be

$$\tilde{\theta}_i = \left(\frac{\tau^{-1}}{\tau^{-1} + \hat{\sigma}_i^{-2}}\right)\mu + \left(\frac{\hat{\sigma}_i^{-2}}{\tau^{-1} + \hat{\sigma}_i^{-2}}\right)\hat{\theta}_i, \quad (7)$$

with an estimated variance

$$\hat{V}(\tilde{\theta}_i) = \frac{1}{\tau^{-1} + \hat{\sigma}_i^{-2}}. \quad (8)$$

The overall effect μ lies in the middle of $\hat{\theta}_1, \dots, \hat{\theta}_N$. The strengthened estimate (7) is a weighted average of the original estimate and the overall effect, with weights determined by their relative precisions. This method shrinks the estimates toward a common value. The standard errors of the estimates also shrink, because the new variance (8) is smaller than $\hat{\sigma}_i$. For these reasons, the method is often called shrinkage estimation [28]. In practice, μ and τ are unknown and need to be estimated, and the uncertainty about those parameters should be reflected in standard errrors; various techniques for doing this are available [26].

Accounting for Correlations Among Categories

In most meta-analyses, the estimates come from different studies and are regarded as independent. In this case, independence is implausible, because many of the crime categories overlap. For example, more than 80% of violent crimes are assaults, so the effects of refresher training on violent crime and assault should be similar. Meta-analyses are not adversely impacted if the studies are mildly correlated, but the overlap between some of our categories is too large to ignore.

Overlap exists both in the sample and the population. For this reason, we devised a multivariate meta-analysis procedure that accounts

Table 7: Estimated intervention effects, standard errors and p-values for categories of personal crime by whether the crime was reported to police

	All crimes			Reported to police			Not reported to police		
	Coef	SE	*p**	Coef	SE	*p**	Coef	SE	*p**
Violent crime[a]	.133	.178	.455	−.291	.237	.219	.589	.287	.031
Serious violent crime[b]	.123	.284	.665	−.247	.341	.471	.482	.558	.387
Rape/sexual assault	−.632	.957	.499	−1.92	2.06	.312	.176	1.28	.881
Robbery	.555	.538	.301	.445	.693	.514	.982	1.08	.343
Assault	.081	.191	.670	−.295	.265	.263	.503	.298	.079
Aggravated	.043	.400	.910	−.187	.494	.682	.521	.955	.578
Simple	.160	.234	.495	−.235	.341	.483	.501	.333	.125
Personal theft[c]	.396	1.62	.772	−.459	1.94	.872	.451	1.72	.808

** equal-tailed Bayesian p-value*

[a] *excludes homicide, because the NCVS is based on interviews with victims and therefore cannot measure murder*

[b] *includes rape or sexual assault, robbery and aggravated assault*

[c] *includes pocket picking, completed purse snatching and attempted purse snatching*

for correlations among the estimated effects and among the true effects. Let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_N)^T$ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)^T$ denote vectors of estimated and true effects. We suppose that $\hat{\boldsymbol{\theta}}$ has a joint normal distribution around $\boldsymbol{\theta}$,

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is an $N \times N$ covariance matrix. We then suppose that $\boldsymbol{\theta}$ is distributed as

$$\boldsymbol{\theta} \sim N(\mu \boldsymbol{1}, \tau \boldsymbol{R}),$$

where $\boldsymbol{1} = (1,\dots,1)^T$ is a vector of ones, τ is a between-study variance, and $\boldsymbol{R}$ is a between-study correlation matrix. Setting the mean of $\boldsymbol{\theta}$ equal to $\mu \boldsymbol{1}$ constrains the true effects to vary around a single overall effect μ.

Shrinkage estimates and standard errors for this model come from multivariate versions of (7)–(8); formulas are provided in the technical appendix.

Specifying the Correlations

Correlations among the true effects are unknown, so we applied rough guesses by the following procedure. We set each element R_{jk} of the matrix $\boldsymbol{R}$ equal to the number of crimes that are common to categories j and k, divided by the total number of crimes in categories j and k combined. For example, consider the two categories of assault and serious violent crime. Assaults are classified as simple or aggravated; the latter are considered serious violent crimes, but the former are not. In the 2011 NCVS, interviewers discovered 1,117 assaults, of which 841 were simple and 276 were aggravated, and they discovered 499 serious violent crimes. The number of crimes common to the two categories (assault and serious violent crime) is 276, and the number of crimes in both categories combined is 1,117 + 499 − 276 = 1,340, so the correlation between these two categories was set to 276 / 1,340 = 0.206.

The correlations among the estimated effects are also unknown. Because we fit a model to each category of crime separately, no direct estimates of these correlations are available. However, it is reasonable to suppose that correlations among the estimated effects and correlations among the true effects will be similar. For this reason, we set each element Σ_{jk} of $\boldsymbol{\Sigma}$ to

$$\Sigma_{jk} = \hat{\sigma}_j \, \hat{\sigma}_k \, R_{jk},$$

where $\hat{\sigma}_j$ and $\hat{\sigma}_k$ are the standard errors for $\hat{\theta}_j$ and $\hat{\theta}_k$ obtained from our models.

Having specified $\boldsymbol{\Sigma}$ and $\boldsymbol{R}$, the remaining quantities to be estimated are the overall effect μ, the between-study variance τ, and the vector of true effects $\boldsymbol{\theta}$. We estimated them jointly by a Bayesian procedure using Markov chain Monte Carlo; details of the method are given in the appendix.

Results

Using the estimated effects and standard errors for all crimes (reported to police or not) from Tables 6 and 7, we applied the meta-analysis procedure to obtain a pooled estimate of the overall training effect and updated estimates for each category of personal and property crime. We then repeated the

Table 8: Estimated intervention effects, standard errors and p-values for categories of personal and property crime resulting from meta-analysis

	All crimes			Reported to police			Not reported to police		
	Coef	SE	p^*	Coef	SE	p^*	Coef	SE	p^*
Overall effect	.212	.121	.090	–.024	.168	.924	.455	.168	.013
Violent crime[a]	.191	.130	.151	–.121	.195	.557	.492	.191	.010
Serious violent crime[b]	.198	.156	.198	–.096	.219	.696	.459	.236	.054
Rape/sexual assault	–.185	.214	.288	–.057	.306	.892	. 446	.274	.092
Robbery	.245	.185	.147	.014	.259	.949	.479	.272	.061
Assault	.174	.137	.212	–.115	.203	.601	.466	.190	.018
Aggravated	.186	.176	.258	–.067	.242	.818	.455	.261	.073
Simple	.199	.143	.166	–.083	.218	.738	.465	.198	.022
Personal theft[c]	.215	.212	.231	–.031	.303	.960	.454	.288	.090
Property crime	.265	.081	.001	.081	.117	.484	.401	.097	.000
Household burglary	.272	.126	.026	.107	.175	.545	.516	.187	.003
Motor vehicle theft	.123	.218	.441	–.129	.271	.685	.415	.280	.119
Theft	.268	.087	.002	.054	.130	.678	.404	.103	.000

* *equal-tailed Bayesian p-value*

[a] *excludes homicide, because the NCVS is based on interviews with victims and therefore cannot measure murder*

[b] *includes rape or sexual assault, robbery and aggravated assault*

[c] *includes pocket picking, completed purse snatching and attempted purse snatching*

meta-analysis for crimes reported to police and for crimes not reported to police. Results from these meta-analyses are shown in Table 8.

Comparing the results in Table 8 to those in Tables 6 and 7, we see that the new estimates are smoother, with less variation across crime categories, and the standard errors have been reduced. Estimates that were least precise have changed the most. For crimes not reported to police, the estimated overall intervention effect (.455) is large and highly significant. The exponentiated value is

$$\exp(0.455) = 1.58,$$

which suggests that the intervention increased the reporting of these crimes by about 58%. For crimes reported to police, the overall effect (–.024) is close to zero and insignificant. For all crimes (reported or not), the overall effect (.212) lies between them and is nearly significant (p = .090).

EFFECTS OF THE INTERVENTION ON DATA QUALITY INDICATORS

Thus far, we have described the effects of the intervention on key survey outcomes related to crime rates. We also analyzed the effects of training on two key indicators of data quality: average screener times and household response rates.

Table 9 shows the average screener interview times in seconds, and the number of persons interviewed, by cohort and time period. (The numbers of interviews differ from those in Table 1, because not all interviews yielded usable screener times or crime counts.)

Table 9: Average screener times in the 2011 NCVS refresher training experiment by cohort and time period

	Cohort 1		Cohort 2	
	Time 1	Time 2	Time 1	Time 2
Average	87.2	172.8	88.2	89.2
Interviews	54,056	25,241	55,391	25,033

With intervention, the average for Cohort 1 rose by 172.8 – 87.2 = 85.6 seconds. Without intervention, the average for Cohort 2 rose by 89.2 – 88.2 = 1.0 seconds. An estimate of the intervention effect on the screener times is the difference of the differences,

$$(172.8 - 87.2) - (89.2 - 88.2) = 84.6.$$

To obtain a proper standard error for this effect, we need to account for the experimental design. In a similar vein as our previous models, let Y_{ijkt} and N_{ijkt} denote the average screener time and number of interviews, respectively, for interviewer i in Cohort j and team pair k during Time t. In this case we will use a weighted normal linear regression model with random effects for interviewers and team pairs,

$$\begin{aligned} Y_{ijkt} &\sim N(\,\mu_{ijkt},\,\sigma^2/N_{ijkt}\,), \\ \mu_{ijkt} &= \boldsymbol{x}_{ijkt}^T\boldsymbol{\beta} + \alpha_i + \boldsymbol{x}_{ijkt}^T\gamma_k. \end{aligned} \tag{9}$$

The vector of predictors $\boldsymbol{x}_{ijkt}$ has the same form as before, with a constant, a dummy indicator for Time = 2, a dummy indicator for Cohort = 1, and a Time × Cohort product. The random effect for interviewer i is distributed as

$$\alpha_i \sim N(0,\,\sigma_\alpha^2),$$

and the team-pair effects are jointly normally distributed,

$$\gamma_k \sim N(0,\,\Sigma_\gamma\,),$$

where Σ_γ is a 4×4 covariance matrix. The procedures for fitting this model are similar to those used for the previous ones. Key results from this model are shown in Table 10.

Table 10: Coefficients, standard errors, and p-values from expanded model for average screener times

	Coef	SE	p^*
Constant	88.17	2.48	—
Time	0.915	4.44	.837
Cohort	−0.972	3.53	.783
Time×Cohort	84.65	6.28	.000

**equal-tailed Bayesian p-value*

We estimate that the intervention raised average screener times by about 85 seconds, and the effect is highly significant.

Table 11 shows the number of households interviewed, the number of households attempted, and the percent household response rate.

Table 11: Response rates in the 2011 NCVS refresher training experiment by cohort and time period

	Cohort 1		Cohort 2	
	Time 1	Time 2	Time 1	Time 2
Interviewed	24,603	11,841	19,781	10,144
Attempted	27,157	13,403	21,709	11,351
Rate*	90.6	88.3	91.1	89.4

**rate per 100 interviews*

For this analysis, we use a binomial logistic model. Let Y_{ijkt} and N_{ijkt} denote the proportion of households successfully interviewed, and the number of households attempted, for interviewer i in Cohort j and team pair k during Time t. We assume

$$\begin{aligned} Y_{ijkt} &\sim N_{ijkt}^{-1}\,\text{Bin}(\,\mu_{ijkt},\,N_{ijkt}\,), \\ \log\left(\frac{\mu_{ijkt}}{1-\mu_{ijkt}}\right) &= \boldsymbol{x}_{ijkt}^T\boldsymbol{\beta} + \alpha_i + \boldsymbol{x}_{ijkt}^T\gamma_k, \\ \alpha_i &\sim N(0,\,\sigma_\alpha^2), \\ \gamma_k &\sim N(0,\,\Sigma_\gamma\,). \end{aligned} \tag{10}$$

Results from this model are shown in Table 12.

Table 12: Coefficients, standard errors, and p-values from expanded model for households response rates

	Coef	SE	p^*
Constant	2.419	0.086	—
Time	−0.202	0.058	.000
Cohort	−0.100	0.103	.377
Time×Cohort	−0.065	0.077	.408

**equal-tailed Bayesian p-value*

The estimated intervention effect of −0.100, which is on the log-odds scale, is small and insignificant. The intervention had no discernible effect on household response rates.

IMPLICATIONS

Using a family of statistical models that accounts for the experimental design, we have estimated the combined effects of refresher training and performance monitoring within major categories of personal and property crime. For many of these categories, crime counts were small and the estimates were imprecise. However, when we borrowed strength across the categories by a meta-analysis, we found a consistent pattern of intervention-related increase in crimes not reported to police, and no discernible training effects among crimes reported to police. We also found that the intervention led to a significant increase in length of the screener interview.

As mentioned in the Background section of this report, the intervention had no effect on any published results from 2011, because all post-intervention interviews from Cohort 1 (about one-eighth of all interviews conducted in 2011) were removed from the sample prior to weighting and estimation.

Because refresher training and performance monitoring were enacted within each cohort at approximately the same time, this experiment does not allow us to separate the effects of these components. These results provide little guidance for strategizing about future interventions. At present, the performance monitoring system based on the new data quality indicators remains in place. New interviewers who join the NCVS workforce receive training that is essentially equivalent to the refresher training described in this document, but currently there are no firm plans to retrain experienced interviewers who received refresher training in 2011 or 2012.

One key question not addressed in this report is whether the intervention effects seen in Cohort 1 in the latter months of 2011 were sustained into 2012. Another key question is what effect, if any, the intervention had on Cohort 2. Answers to those questions will determine whether and how the results from NCVS 2012 may be compared to those in 2011 and previous years. Analyses from a companion report by Schafer (2013) suggest that the combined effects of the intervention in Cohorts 1 and 2 on the reporting of crime in 2012 were small and not statistically significant [9].

ACKNOWLEDGMENTS

William Sabol and Michael Planty at BJS helped to guide this research, and Meagan Wilson supplied an overview of refresher training and performance monitoring programs. Many useful suggestions were provided by William Samples, Patrick Flanagan, Rod Little, Stephen Ash, Derek Young, and members of the NCVS Data Review Panel. Special thanks to David Watt and Lindsay Longsine for creating and managing the required datasets.

REFERENCES

[1] Biderman, A.D.D., Cantor, D., Lynch, J.P. and Martin, E. (1986) *Final Report of Research and Development for the Redesign of the National Crime Survey.* Prepared for the Bureau of Justice Statistics. Washington, DC: Bureau of Social Science Research, Inc.

[2] Billiet, J. and Loosveldt, G. (1988) Improvement of the quality of responses to factual survey questions by interviewer training. *Public Opinion Quarterly*, 52. 190–211.

[3] Fowler, F.J. and Mangione, T.W. (1990) *Standardized Survey Interviewing.* Thousand Oaks, CA: Sage Publications.

[4] Groves, R.M. and McGonagle, K.A. (2001) A theory-guided interviewer training protocol regarding survey participation. *Journal of Official Statistics*, 17, 249–265.

[5] Dahlhamer, J.M., Cynamon, M.L., Gentleman, J.F., Piani, A.L. and Weiler, M.J. (2010) Minimizing survey error through interviewer training: new procedures applied to the National Health Interview Survey. *Section on Survey Research Methods*, JSM 2010, American Statistical Association.

[6] Miller, P.V. and Groves, R.M. (1985) Matching survey responses to official records: An exploration of validity in victim reporting. *The Public Opinion Quarterly*, 49, 366–380.

[7] Czaja, R., Blair, J., Bickart, B., Eastman, E. (1994) Respondent strategies for recall of crime victimization incidents. *Journal of Official Statistics*, 10, 257–276.

[8] Planty, M. and Truman, J.L. (2012) *Criminal Victimization, 2011.* U.S. Department of Justice, Bureau of Justice Statistics, NCJ 239437, October 2012.

[9] Schafer, J.L. (2013) Modeling the effects of field interventions in the 2012 National Crime Victimization Survey. Working paper dated June, 2013 prepared for the NCVS Data Review Panel. Center for Statistical Research and Methodology, U.S. Census Bureau.

[10] van den Brakel, J.A. (2008) Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey. *Journal of the Royal Statistical Society Series A*, 171, 581–613.

[11] Murray, D.M. (1998) *Design and Analysis of Community Trials.* Oxford, UK: Oxford University Press.

[12] Donner, A. and Klar, N. (2000) *Design and Analysis of Cluster Randomization Trials in Health Research.* New York: Oxford University Press.

[13] The merits of matching in community intervention trials: A cautionary tale. *Statistics in Medicine*, 16, 1753–1764.

[14] Imai, K., King, G. and Nall, C. (2009) The essential role of pair matching on cluster-randomized experiments, with

application to the Mexican Universal Health Insurance Evaluation. *Statistical Science*, 24, 29–53.

[15] Randomization by group: A formal analysis. *American Journal of Epidemiology*, 108, 100–102.

[16] Hill, J. and Scott, M. (2009) Comment: The essential role of pair matching. *Statistical Science*, 24, 54–58.

[17] Truman, J.L. (2011) *Criminal Victimization, 2010*. U.S. Department of Justice, Bureau of Justice Statistics, NCJ 235508, September 2011.

[18] McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, Second edition. New York: Chapman & Hall.

[19] Agresti, A. (2002) *Categorical Data Analysis*, Second edition. New York: Wiley.

[20] R Development Core Team (2011) *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

[21] Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9–25.

[22] Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004) *Applied Longitudinal Analysis.* Hoboken, NJ: Wiley.

[23] Rabe-Hesketh, S. and Skrondal, A. (2008) *Multilevel and Longitudinal Modeling Using Stata*, Second edition. College Station, TX: Stata Press.

[24] Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2009) *MLwiN Version 2.1.* Centre for Multilevel Modelling, University of Bristol.

[25] Gelman, A. and Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.

[26] Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A. and Song, F. (2000) *Methods for Meta-Analysis in Medical Research.* Chicester, West Sussex: Wiley.

[27] DerSimonian, R. and Laird, N. (1986) Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7, 177–188.

[28] Carlin, B.P. and Louis, T.A. (2009) *Bayesian Methods for Data Analysis*, Third Edition. Boca Raton, FL: Chapman and Hall/CRC Press.

[29] Gelman, A., Rubin, D.B., Carlin, J., and Stern, H. (2004) *Bayesian Data Analysis*, Second edition. London: Chapman & Hall/CRC Press.

[30] Liang, F., Liu, C. and Carroll, R.J. (2010) *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples.* New York: Wiley.

[31] Brooks, S., Gelman, A., Jones, G. and Meng, X.-L. (2011) *Handbook of Markov Chain Monte Carlo*, New York: Chapman & Hall/CRC Press.

[32] Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.

[33] Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, 22, 1701–1762.

TECHNICAL APPENDIX

Bayesian Model Formulation

Our model for training effects is a nonlinear mixed-effects regression with three levels of nested observations, and most programs for mixed-effects modeling are unable to handle it. Likelihood-based methods are prone to fail in this situation, because crimes are rare events and the data are sparse. For these reasons, we adopted a fully Bayesian approach, simulating parameter estimates by Markov chain Monte Carlo (MCMC). Bayesian methods for data analysis are described by Gelman et al. (2004) [29] and by Carlin and Louis (2009) [28]; for a thorough, applied treatment of Bayesian multilevel modeling, see Gelman and Hill (2007) [25]. Here we describe the computational procedures for the Poisson version of the model; procedures for the normal and binomial versions are minor variants of this.

In our model, we assumed that the distribution for Y_{ijkt} given $\boldsymbol{\beta}$, α_i and γ_k is

$$Y_{ijkt} \,|\, \boldsymbol{\beta}, \alpha_i, \gamma_k \sim \text{Poisson}(\mu_{ijkt}),$$

where

$$\log \mu_{ijkt} = \log N_{ijkt} + \boldsymbol{x}_{ijkt}^T \boldsymbol{\beta} + \alpha_i + \boldsymbol{x}_{ijkt}^T \gamma_k.$$

We assumed that the interviewer effects were distributed as

$$\alpha_i \,|\, \sigma_\alpha^2 \sim N(0, \sigma_\alpha^2)$$

independently for all interviewers, and that the team-pair effects were distributed as

$$\gamma_k \,|\, \mathbf{\Sigma}_\gamma \sim N(0,\, \mathbf{\Sigma}_\gamma)$$

independently for all team pairs. Finally, we applied prior distributions to $\boldsymbol{\beta}$, σ^2_α and $\mathbf{\Sigma}_\gamma$. Following standard practice, we used an improper uniform density for $\boldsymbol{\beta}$, which can be regarded as the limiting form of a multivariate normal density $N(0,\, \mathbf{\Sigma}_\beta)$ as $\mathbf{\Sigma}^{-1}_\beta \rightarrow 0$. For the interviewer variance, we used a scaled inverted chisquare distribution with scale factor a and degrees of freedom b,

$$\sigma^2_\alpha \sim a\,\chi^{-2}_b. \tag{11}$$

For the team-pair covariance matrix, we used an inverted Wishart distribution

$$\mathbf{\Sigma}^{-1}_\gamma \sim \text{Wishart}(c, \boldsymbol{D}), \tag{12}$$

where c is the degrees of freedom and $\boldsymbol{D}$ is the scale matrix. For the hyperparameters, we chose $a = 1$, $b = 1$, $c = \dim(\gamma_k)$ and $\boldsymbol{D} = c^{-1}\boldsymbol{I}$. These priors are diffuse, reflecting vague knowledge about the variance components with rough prior guesses $\sigma^2_\alpha \approx 1$ and $\mathbf{\Sigma}_\gamma \approx \boldsymbol{I}$.

Blocked Gibbs Sampler

To describe the MCMC procedure, we need some additional notation. Let $\boldsymbol{Y} = \{Y_{ijkt}\}$ denote the set of observed responses, and let

$$\mathbf{\Theta} = \{\boldsymbol{\beta},\, \{\alpha_i\},\, \{\gamma_k\},\, \sigma^2_\alpha,\, \mathbf{\Sigma}_\gamma\}$$

denote all the unknown quantities in our model. Let $\setminus$ denote the relative complement set operator, so that

$$\mathbf{\Theta} \setminus \boldsymbol{\beta} = \{\{\alpha_i\},\, \{\gamma_k\},\, \sigma^2_\alpha,\, \mathbf{\Sigma}_\gamma\}$$

contains all components of $\mathbf{\Theta}$ except $\boldsymbol{\beta}$. Finally, let square brackets denote a distribution, so that

$$[\boldsymbol{\beta} \mid \boldsymbol{Y}, \mathbf{\Theta} \setminus \boldsymbol{\beta}]$$

is the conditional posterior distribution for $\boldsymbol{\beta}$ given $\boldsymbol{Y}$ and all other components of $\mathbf{\Theta}$.

We simulated draws of $\mathbf{\Theta}$ from the joint posterior distribution $[\mathbf{\Theta} \,|\, \boldsymbol{Y}]$ using a blocked Metropolis-within-Gibbs strategy [30] [31]. Suppose that we could draw from the conditional distributions

$$\begin{aligned} \boldsymbol{\beta} &\sim [\boldsymbol{\beta} \mid \boldsymbol{Y}, \mathbf{\Theta} \setminus \boldsymbol{\beta}], \\ \alpha_i &\sim [\alpha_i \mid \boldsymbol{Y}, \mathbf{\Theta} \setminus \alpha_i], \\ \gamma_k &\sim [\gamma_k \mid \boldsymbol{Y}, \mathbf{\Theta} \setminus \gamma_k], \\ \sigma^2_\alpha &\sim [\sigma^2_\alpha \mid \boldsymbol{Y}, \mathbf{\Theta} \setminus \sigma^2_\alpha], \\ \mathbf{\Sigma}_\gamma &\sim [\mathbf{\Sigma}_\gamma \mid \boldsymbol{Y}, \mathbf{\Theta} \setminus \mathbf{\Sigma}_\gamma]. \end{aligned}$$

Repeating this cycle many times would eventually produce a draw from $[\mathbf{\Theta} \,|\, \boldsymbol{Y}]$. The technique of sampling from the full conditional distribution for each component given the other components is called a Gibbs sampler. In a classic Gibbs sampler, each of the simulated components is one-dimensional. If some components are multidimensional, the Gibbs sampler is said to be blocked. Blocking generally leads to faster convergence, meaning that fewer cycles are needed to approximate the stationary distribution, sometimes at the cost of greater computational complexity per cycle. In this particular application, the blocking is natural and convenient.

In the blocked Gibbs sampler, the conditional distribution for the interviewer variance is

$$[\sigma^2_\alpha \mid \boldsymbol{Y}, \mathbf{\Theta} \setminus \sigma^2_\alpha] = a'\,\chi^{-2}_{b'},$$

where the updated hyperparameters are

$$\begin{aligned} a' &= a + \sum_{i=1}^{n_I} \alpha_i^2, \\ b' &= b + n_i, \end{aligned}$$

and n_i is the number of interviewers. The conditional distribution for the team-pair covariance matrix is

$$[\mathbf{\Sigma}^{-1}_\gamma \mid \boldsymbol{Y}, \mathbf{\Theta} \setminus \mathbf{\Sigma}_\gamma] = \text{Wishart}(c', \boldsymbol{D}'),$$

where

$$\begin{aligned} c' &= c + n_k, \\ \boldsymbol{D}' &= \boldsymbol{D} + \sum_{k=1}^{n_k} \gamma_k \gamma_k^T, \end{aligned}$$

and n_k is the number of team pairs. These two distributions are straightforward, but the other three in the blocked Gibbs sampler are not. The conditional posterior distributions for $\boldsymbol{\beta}$, α_i and γ_k are nonstandard, and producing exact draws from them would be difficult. Following standard practice, we replaced the exact simulation of each of these conditional distributions by one step of a Metropolis-Hastings algorithm that converges to the corresponding conditional.

Metropolis-Hastings

Consider the problem of simulating draws of a random vector $\boldsymbol{\theta}$ whose probability density function is $f(\boldsymbol{\theta})$, which is called a target density. Metropolis-Hastings (MH) proceeds as follows [32]. Let $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ be the state of the process at iteration t. We need a jumping rule, which is often called a proposal density, to generate a candidate value $\boldsymbol{\theta}^*$. This proposal may depend on $\boldsymbol{\theta}^{(t)}$, so we write the proposal density as $q(\boldsymbol{\theta}^* \,|\, \boldsymbol{\theta}^{(t)})$. After generating $\boldsymbol{\theta}^*$ from the proposal, we compute the MH acceptance ratio

$$r(\boldsymbol{\theta}^{(t)} \,|\, \boldsymbol{\theta}^*) \;=\; \frac{f(\boldsymbol{\theta}^*) \,/\, f(\boldsymbol{\theta}^{(t)})}{q(\boldsymbol{\theta}^* \,|\, \boldsymbol{\theta}^{(t)}) \,/\, q(\boldsymbol{\theta}^{(t)} \,|\, \boldsymbol{\theta}^*)}.$$

We then generate a standard uniform random variate $U \sim U(0,1)$ and take

$$\boldsymbol{\theta}^{(t+1)} \;=\; \begin{cases} \boldsymbol{\theta}^* & \text{if } U \leq r(\boldsymbol{\theta}^{(t)} \,|\, \boldsymbol{\theta}^*)\,, \\ \boldsymbol{\theta}^{(t)} & \text{otherwise.} \end{cases}$$

If f includes an intractable normalizing constant, that constant will drop out of the MH ratio. Any factor in $q(\boldsymbol{\theta}^* \,|\, \boldsymbol{\theta}^{(t)})$ that does not depend on $\boldsymbol{\theta}^{(t)}$ will also drop out.

There are two common strategies for choosing a proposal for MH. One is to use an ellipsoidal distribution centered at $\boldsymbol{\theta}^{(t)}$, for example,

$$\boldsymbol{\theta}^* \;=\; \boldsymbol{\theta}^{(t)} \,+\, \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$ for some covariance matrix $\boldsymbol{\Sigma}$. The other common strategy is to choose a proposal that is intended to closely approximate the target. For example, we might use

$$\boldsymbol{\theta}^* \;=\; \tilde{\boldsymbol{\theta}} \,+\, \boldsymbol{\epsilon},$$

where $\tilde{\boldsymbol{\theta}}$ is the mode of the target density $f(\boldsymbol{\theta})$, and $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$ for some $\boldsymbol{\Sigma}$ that approximates the covariance matrix of $\boldsymbol{\theta}$. This proposal does not depend on the current state $\boldsymbol{\theta}^{(t)}$, and the resulting algorithm is called an independence sampler [33]. Independence samplers are prone to getting stuck, which often happens when the target density has heavier tails than the proposal. Switching the proposal from a multivariate normal to a multivariate t with small degrees of freedom will often solve the problem.

For this application, we embedded MH algorithms into the blocked Gibbs sampler to replace the intractable conditional distributions for β, α_i and γ_k. Each of those intractable conditionals has essentially the same form: the posterior distribution from a Poisson loglinear regression with a multivariate normal prior for the coefficients. For example, consider the conditional distribution for γ_k,

$$\gamma_k \sim [\,\gamma_k \mid \boldsymbol{Y}, \Theta \setminus \gamma_k\,].$$

This can be viewed as the posterior distribution from the model

$$\begin{aligned} Y_{ijkt} &\sim \text{Poisson}(\mu_{ijkt}), \\ \log \mu_{ijkt} &= \omega_{ijkt} + \boldsymbol{x}_{ijkt}^T \gamma_k, \end{aligned}$$

where

$$\omega_{ijkt} \;=\; \log N_{ijkt} \,+\, \boldsymbol{x}_{ijkt}^T \beta \,+\, \alpha_i$$

is a known offset term, and $\gamma_k \sim N(0, \boldsymbol{\Sigma}_\gamma)$ is a prior distribution for the coefficients. Using Bayes' Theorem, the posterior density for γ_k is, except for an intractable normalizing constant,

$$\begin{aligned} f(\gamma_k) \;\propto\; & \exp\left\{ -\frac{1}{2}\, \gamma_k^T \boldsymbol{\Sigma}_\gamma^{-1} \gamma_k \right\} \\ & \times \exp \sum_{ijt} \left\{ Y_{ijkt} \log \mu_{ijkt} \,-\, \mu_{ijkt} \right\}, \end{aligned}$$

where the sum is taken over all units i, j and t within team pair k. Our proposal distribution is a multivariate t with 4 degrees of freedom, whose center is the mode of f, and whose scale is chosen to match the second derivatives of $\log f$. The mode of f is computed by a Newton-Raphson procedure, making use of the gradient

$$\frac{\partial}{\partial \gamma_k} \log f \;=\; -\,\boldsymbol{\Sigma}_\gamma^{-1} \gamma_k \,+\, \sum_{ijt} \boldsymbol{x}_{ijkt} (y_{ijkt} - \mu_{ijkt})$$

and the Hessian

$$-\,\frac{\partial^2}{\partial \gamma_k \partial \gamma_k^T} \log f \;=\; \boldsymbol{\Sigma}_\gamma^{-1} \,+\, \sum_{ijt} \mu_{ijkt} \boldsymbol{x}_{ijkt} \boldsymbol{x}_{ijkt}^T.$$

Implementation

The Metropolis-within-Gibbs procedure was implemented in Fortran 95 routines called from R. Examining the output stream from exploratory runs, we found that autocorrelations in the key parameters died down very quickly. To produce the estimates for each category of crime shown in Tables 5–7, we ran the algorithm for 10,100 cycles, treating the first 100 as a burn-in period and discarding their results. Averages of the remaining 10,000 iterates provided the estimated coefficients, and standard deviations of the iterates provided the standard errors. Quantiles of the iterates were used to compute posterior intervals, and the Bayesian p-values were defined as one minus the probability content of the widest equal-tailed posterior interval that failed to cover the null value.

Meta-Analysis

For the meta-analysis, let $\hat{\theta}$ and θ denote the vectors of estimated and true effects, respectively. If we suppose that

$$\begin{aligned} \hat{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} &\sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}), \\ \boldsymbol{\theta} \,|\, \mu, \tau &\sim N(\mu \boldsymbol{1}, \tau \boldsymbol{R}), \end{aligned}$$

with Σ and $\boldsymbol{R}$ regarded as known, then the posterior distribution for $\boldsymbol{\theta}$ given $\hat{\boldsymbol{\theta}}$, μ and τ is multivariate normal with mean

$$\begin{aligned} E(\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\theta}},\mu,\tau) &= \left(\tau^{-1}\boldsymbol{R}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1} \\ &\quad \times \left(\tau^{-1}\boldsymbol{R}^{-1}\mu\boldsymbol{1} + \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}}\right) \end{aligned}$$

and covariance matrix

$$V(\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\theta}},\mu,\tau) = \left(\tau^{-1}\boldsymbol{R}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}.$$

These are natural multivariate extensions of the shrinkage formulas (7) and (8). To estimate μ and τ along with $\boldsymbol{\theta}$, we applied improper uniform prior densities to μ and τ and implemented the blocked Gibbs sampler

$$\begin{aligned} \boldsymbol{\theta} &\sim [\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\theta}},\,\mu,\,\tau], \\ \mu &\sim [\mu\,|\,\hat{\boldsymbol{\theta}},\,\boldsymbol{\theta},\,\tau], \\ \tau &\sim [\tau\,|\,\hat{\boldsymbol{\theta}},\,\boldsymbol{\theta},\,\mu]. \end{aligned}$$

The first of these conditional distributions is multivariate normal with mean and covariance matrix given above. The second is univariate normal with mean

$$E(\mu\,|\,\hat{\boldsymbol{\theta}},\,\boldsymbol{\theta},\,\tau) = \left(\boldsymbol{1}^T\boldsymbol{R}^{-1}\boldsymbol{1}\right)^{-1}\,\boldsymbol{1}^T\boldsymbol{R}^{-1}\hat{\boldsymbol{\theta}}$$

and variance

$$V(\mu\,|\,\hat{\boldsymbol{\theta}},\,\boldsymbol{\theta},\,\tau) = \tau\left(\boldsymbol{1}^T\boldsymbol{R}^{-1}\boldsymbol{1}\right)^{-1}.$$

The third is scaled inverted chisquare,

$$[\tau\,|\,\hat{\boldsymbol{\theta}},\,\boldsymbol{\theta},\,\mu] = (\boldsymbol{\theta} - \mu\boldsymbol{1})^T\boldsymbol{R}^{-1}(\boldsymbol{\theta} - \mu\boldsymbol{1})\,\chi^{-2}_{N-2},$$

where $N = \dim(\boldsymbol{\theta})$. To generate the results in Table 8, we implemented this algorithm in R, running it for 50,100 cycles and discarding the first 100 as a burn-in period.