RESEARCH REPORT SERIES
(Statistics #2014-02)

Modeling the Effects of Recent Field Interventions in the 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: April 3, 2014

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.

Modeling the Effects of Recent Field Interventions in the National Crime Victimization Survey

Research Report Series (Statistics)

Joseph L. Schafer

March 2014

RRS2014-02

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.

HIGHLIGHTS

NCVS field staff experienced two major interventions in 2012: completion of a refresher training and performance monitoring program which began the previous year, and a field realignment effort which reduced the number of Census Bureau Regional Offices from twelve to six. To assess the impact of these interventions, we fit Bayesian longitudinal models describing key quality indicators and survey outcomes over a five-year period (2008–2012). After accounting for long-term trends, annual periodic cycles, characteristics of the interviewers' monthly assignments, and random interviewer variation, we detected some statistically significant effects of the interventions on the following variables.

- *Response rates:* Refresher training and performance monitoring were associated with a modest but significant decrease in household response rates in 2011 but not in 2012.
- *Screener times:* Refresher training and performance monitoring produced large and significant increases in the average duration of the screener interview in 2011 and 2012.
- *Personal crime and household property crime:* Neither the refresher training and performance monitoring program nor field realignment were associated with any significant changes in the collection of crimes in 2012.

Because no significant effects on the collection of crimes were detected, there is no evidence to suggest that victimization rates for 2012 or comparisons between 2012 and previous years were impacted.

1 BACKGROUND

The National Crime Victimization Survey (NCVS) is a household-based demographic survey that yields annual national estimates of property crime and nonfatal violent crime, both reported and not reported to the police. Interviews are conducted year-round by Census Bureau Field Representatives (FRs) through personal visit or by telephone, using a Computer-Assisted Personal Interview (CAPI) instrument. In sampled households, persons of age 12 and older are given screener interviews to determine if they have been victimized during the previous six months. If a crime is reported during the screener, the interview continues with an incident report to ascertain details of the event.

In 2012, NCVS field staff experienced two major interventions that may have affected data quality.

First, a program of refresher training and enhanced performance monitoring, which had begun in late 2011, was continued and completed in 2012. This training reoriented FRs to the purpose of the NCVS and reinforced the importance of following correct procedures, especially during screener interviews. After the training, supervisors began to monitor FR performance using an expanded set of data quality indicators. This program was phased in by an experiment. Teams of FRs were randomly assigned to two cohorts. One cohort received the program in late 2011, and the other received it in early 2012. More details of the program and the randomization procedure are given in Section 2. By the beginning of the second quarter of 2012, over 98% of the experienced NCVS interviewers had

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

completed the training, and the performance monitoring system was in place for all FRs. For the remainder of the year, newly hired FRs were trained as they entered the NCVS workforce and put on the enhanced monitoring program.

Second, field operations for NCVS and all other Census Bureau surveys were consolidated from twelve Regional Offices (ROs) to six. This so-called field realignment was phased in during 2012. By the end of the year, six ROs had closed, and field staff from these closing ROs were reassigned.

In this report, we present new longitudinal analyses of quality indicators and survey outcomes. The data come from approximately 1,900 FRs, 420,000 attempted household interviews and 750,000 personal contacts over the five-year period from 2008 to 2012. Our immediate goal is to characterize the effects of refresher training, performance monitoring and field realignment, to inform us of any potential impact of these interventions on victimization estimates for 2012 and comparisons to previous years. More broadly, these models provide a new methodological framework for understanding temporal patterns and trends in the NCVS and other surveys.

Our analyses focus on two key measures of data quality (household response rate, average duration of the NCVS screener interview) and two key survey outcomes (incident rates for personal crimes and household property crimes). In Section 2, we review the major field interventions that took place in recent years and present graphical summaries of how key data-quality and outcome variables have changed over time. In Section 3, we describe a class of Bayesian generalized linear mixed models with special features that capture the time-varying aspects of the interventions and the response variables. Results for the four outcomes are presented in Sections 4–7, respectively, followed by discussion of the implications in Section 8.

2 RECENT MAJOR INTERVENTIONS AND TRENDS

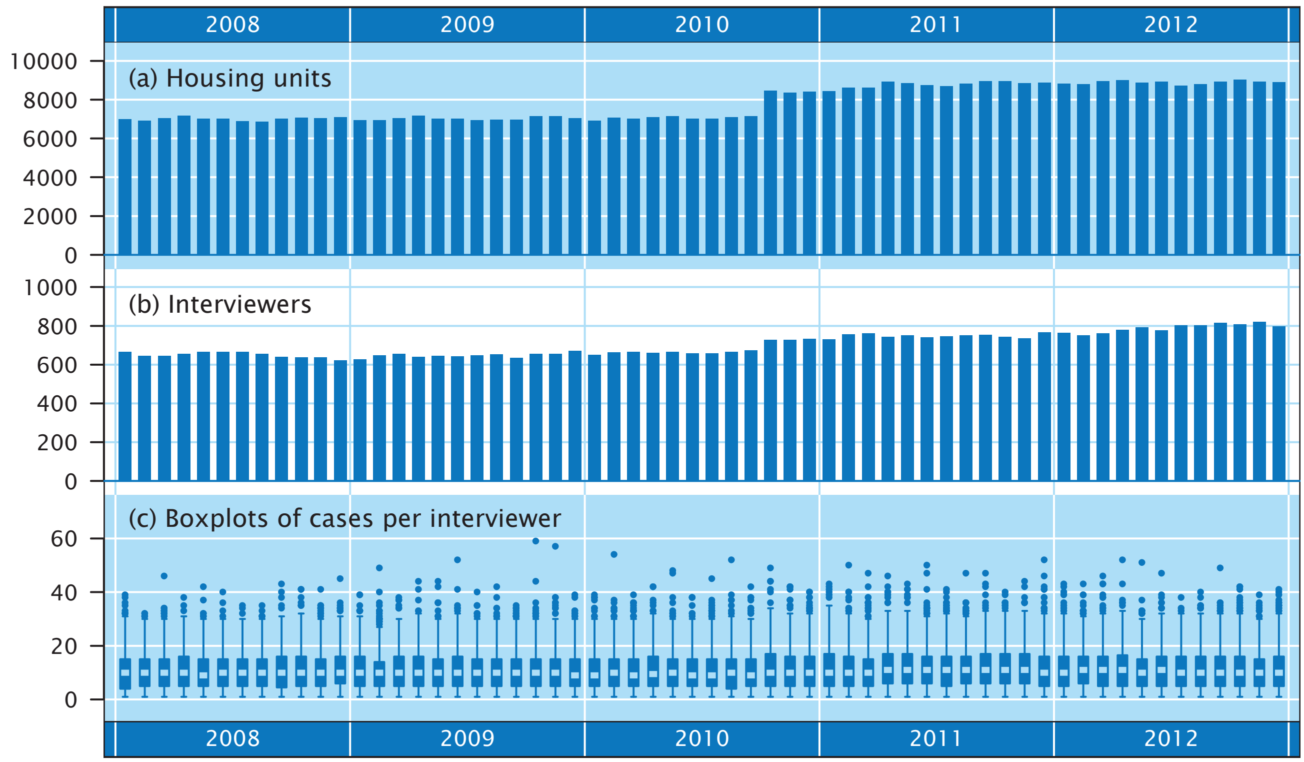
Sample Size and Interviewer Workload

The NCVS, and other major demographic surveys conducted by the Census Bureau, uses a complex two-stage design. The first stage selects Primary Sample Units (PSUs), which are single counties or groups of counties, and the second stage selects housing units and group quarters within the PSUs. Although the same basic design has been used since 2006, the size of the survey has changed over time. Beginning in October 2010, the sample size was boosted by about 25%, reversing some reductions that had been made three years earlier. During this so-called sample reinstatement, the number of interviewers was also increased.

A plot of the number of housing units selected for the NCVS each month from 2008 to 2012 is shown in Figure 1 (a). These figures include units that were successfully interviewed and those that were not, but excludes those that were determined to be out of scope (e.g., because they were vacant or not valid residential addresses). Prior to the sample reinstatement in late 2010, the sample size remained steady at about 7,000 units per month, and then rose to nearly 9,000 by the second quarter of 2011. The number of interviewers working for the NCVS rose over the same period, as shown in Figure 1 (b). The increases in sample size and staffing levels nearly offset each other, and the per-interviewer workload remained fairly steady over time. The distribution of workload (cases per interviewer per month) at each month is shown by the boxplots in Figure 1 (c). The median number of cases per month (represented by each boxplot's center line), and the 25th and 75th percentiles (represented by the edges of the boxes), show remarkably little variation over the five-year period.

In this report, we do not attempt to model the effects of sample reinstatement; those effects, if present, are subsumed into long-term trends and noise. Nor do we attempt to model the impact of the minor changes in interviewer workload over time. However, the wide variation in workload within any given month suggests a diversity in FR experiences which may partly explain some variables of interest. FRs with the highest workloads (up to 60 cases per month) tend to be highly experienced interviewers who work solely on the NCVS. FRs with smaller workloads (as few as one case per month) may represent part-time employees, or they may work primarily on other Census Bureau surveys and receive NCVS cases only sporadically. Each of the models that we will fit include workload as a covariate. Some of our models will also include it as a denominator for a rate or as a precision (inverse-variance) weight. Note that this measure of workload pertains only to NCVS; it does not account for work done by the FR for any other Census Bureau surveys. Workload for other surveys could not be included in our models because it was not available for the entire five-year period.

Figure 1
Monthly Sample Size and Interviewer Workload
(a) Number of housing units selected, (b) number of interviewers working, and (c) boxplots of the distribution of per-interviewer workload (housing units per interviewer) each month



Source: United States Census Bureau, *National Crime Victimization Survey*, 2008–2012

Refresher Training and Performance Monitoring

NCVS refresher training, carried out in the Regional Offices, was a two-day seminar to increase awareness among FRs of the purpose of the survey and the necessity of maintaining high standards of performance. Part of this training was devoted to the screener interview. The screener interview is the portion of the survey where the FR queries the respondent to determine if he or she had been a victim of crime during the previous six months. The screener questions are designed to jog the respondent's memory, helping them to recall crimes that they might otherwise have have overlooked. If the screener is done quickly or haphazardly, incidents of crime could be missed, causing published crime rates to be artificially low. The training sessions also covered several other topics, including

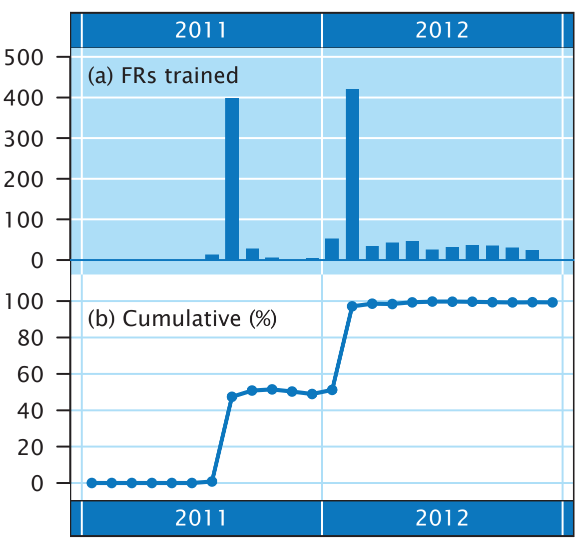
- procedures for completing the NCVS crime incident report,
- use of the Census Bureau's Contact History Instrument (CHI), a system for capturing information about the data collection process, and
- introduction of new data quality field indicators used to measure FR performance.

Before the new field indicators were introduced, FRs were evaluated on the basis of their response rates. The new indicators included measures of household and person-within-household response rates, completeness of the screener questionnaires and crime incident reports, items that had to be changed during the editing and coding process, completeness of the CHI records, duration of the screener and crime incident interviews, and interviews for which the first contact took place outside the monthly data collection period or between the hours of 10 pm and 7 am.

If refresher training and performance monitoring had been applied to all FRs simultaneously, the effects of this intervention would have been difficult to measure, because those effects would have been confounded with changes that occurred for other reasons at the same time. For this reason, the program was phased in through a randomized experiment. Teams of FRs were randomly assigned to two groups. The first group (Cohort 1) received the intervention in 2011, and the second group (Cohort 2) received it in 2012. Applying the intervention to FR teams rather than individual FRs reduced the possibility of Cohort 2 being influenced by the treatment given to Cohort 1, which would have contaminated between-group comparisons. Most of the FRs in Cohort 1 were trained in August 2011, and most of the FRs in Cohort 2 were trained in February 2012, but some who were unable to attend at those times were trained later. By the end of the first quarter of 2012, 98% of the experienced FRs had been trained, and most of the training that took place after that represents new hires who were trained as they joined the workforce. The number of FRs trained each month, and the cumulative percentage of FRs working on the survey who had been trained by each month, are plotted in Figure 2.

In a companion report by Schafer (2013), we modeled the effects of refresher training and performance monitoring on rates of reported crime in 2011 [1]. For the latter months of 2011 when Cohort 1 had been trained but Cohort 2 had not, the training effects were estimated by a between-cohort comparison, and the randomization ensured that the comparison was fair.

Figure 2
Refresher Training by Month
(a) Number of Field Representatives (FRs) trained each month, and (b) cumulative percentage of working FRs who had been trained by each month



Source: United States Census Bureau, *National Crime Victimization Survey*, 2011–2012

As we move into 2012, however, estimating training effects becomes more challenging, because by the end of the first quarter nearly all the FRs had been trained; we no longer have a large control group of untrained FRs to serve as a baseline for comparison.

To estimate training effects in 2012, we lean heavily on the longitudinal nature of these data and on the variation in training dates across interviewers. As interviewers change their status from untrained to trained, any sudden or unexpected shifts in outcomes that cannot be accounted for by other evidence — by long-term trends, by seasonal effects that were seen in previous years, by changes in other covariates, and by random month-to-month variation — will be attributed to training. The scientific basis for inferring causal effects of training is weaker for 2012 than for 2011; our conclusions are more correlational than causal. These inferences require intelligent models that describe how outcome measures evolve under ordinary circumstances when no intervention is taking place, so that when the intervention does takes effect, we have a yardstick for judging whether the observed change is unusual.

Field Realignment

Field realignment was a major restructuring of Census Bureau field operations to reduce the costs of data collection. During realignment, the geographical boundaries covered by the Regional Offices were reconfigured. Six of the physical Regional Offices were closed, and FRs were assigned to a new management structure in the six offices that remained open.

Realignment was phased in throughout the 2012 calendar year. For each of the six non-closing Regional Offices, the new geographic region covered by that office was divided into eight areas. The areas were assigned to seven waves: one area in each of the Waves 1, …, 6 and two areas in Wave 7. The waves transitioned to the new management structure at different dates, beginning with Wave 1 on January 1 and concluding with Wave 7 on November 1. The areas were not assigned to waves in a randomized fashion. However, efforts were made to ensure that the areas within the waves were reasonably well balanced with respect to size and important demographic characteristics.

As field realignment took effect in each wave, interviewers for the NCVS transitioned to the new management structure. The number of FRs transitioning each month, and the cumulative percentage of FRs who had transitioned by each month, are plotted in Figure 3.

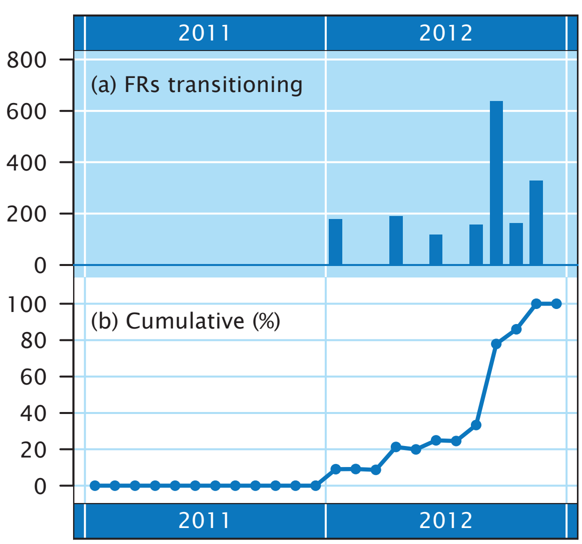
To estimate the effects of field realignment, we will follow a similar strategy as the one we outlined for refresher training and performance monitoring. As interviewers switch over to the new management structure, any sudden shift in outcomes that cannot be explained by long-term trends, seasonal effects, changes in other covariates, and random monthly variation will be attributed to realignment. The validity of this approach will depend on the veracity of the model, and its ability to describe how the outcomes evolve under ordinary circumstances without the intervention.

Two Quality Measures

To assess the effects of refresher training, performance monitoring and field realignment, we will model two variables that are related to data quality.

The first quality measure is the NCVS household response rate. This rate is defined as the number of successful household interviews divided by the number of sampled households, excluding the units determined to be out of scope. A plot of the response rate by month for the period 2008–2012

Figure 3
Field Realignment by Month
(a) Number of Field Representatives (FRs) transitioning each month, and (b) cumulative percentage of working FRs who had transitioned by each month



Source: United States Census Bureau, *National Crime Victimization Survey*, 2011–2012

is shown in Figure 4 (a). The response rates were slowly increasing from 2008 until the first quarter of 2010, and slowly decreasing thereafter. This pattern may be partly explained by the 2010 Decennial Census. The highly visible campaign of public outreach to encourage response to the Decennial Census appears to have had the residual effect of increasing participation in the NCVS; this increase around Census Day (April 1, 2010) and subsequent decline has been seen in other Census Bureau household surveys as well. Careful inspection of this plot also suggests the possibility of annual seasonal effects, such as a dip in response rate at the end of each year during the Christmas holiday season. Thus, our efforts to measure the effects of refresher training and field realignment take place against a backdrop of response rates that have been steadily declining for nearly three years.

The other quality measure is the duration of the screener interview. A plot of the average monthly screener time is shown in Figure 4 (b). This variable sharply increased during late 2011 and early 2012, during the period of refresher training and enhanced performance monitoring. We believe this happened for two reasons. First, a part of the training seminar was specifically devoted to the screener interview, to teaching FRs the importance of strictly following the protocol of askng all of the screener questions, even though many of those questions seem redundant. Second, the training seminar introduced new policies by which the managers were expected to monitor the performance of their field staff. Until then, FRs had been graded solely on their response rates. Screener times had averaged less than 90 seconds, and many screeners were over in less than one minute, suggesting that many FRs were not administering the screener as designed. After training, with the implementation of enhanced monitoring, managers were instructed to use additional quality measures, including screener times, as performance standards. A new benchmark was set for screener times to be at least 3.5 minutes (210 seconds). An increase in screener time was essentially mandated by a change in management policy. Because the training of FRs and the policy change occurred at approximately the same time, it is difficult to separate the effect of FR training from the effect of the policy change.

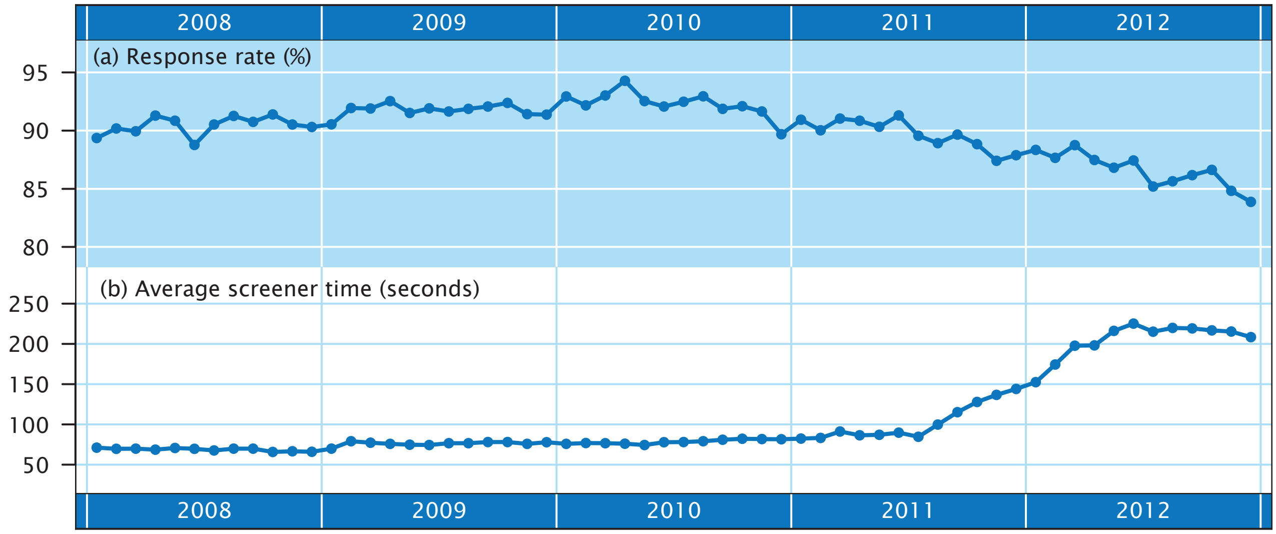
Two Survey Outcomes

In the sections ahead, we will also examine the effects of refresher training and field realignment on two key survey outcomes.

Figure 4

Two Key Measures of Data Quality by Month

(a) Percent household response rate each month, and (b) average duration of the screener interview each month

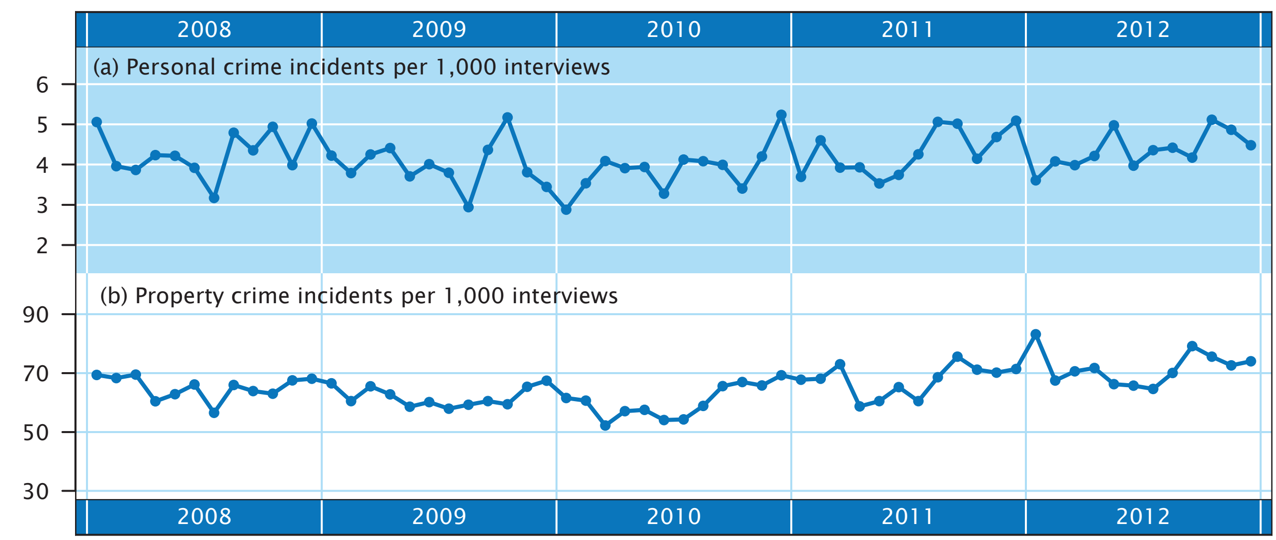


Source: United States Census Bureau, *National Crime Victimization Survey*, 2008–2012

Figure 5

Two Key Survey Outcomes by Month

(a) Incidents of personal crime per 1,000 persons interviewed each month, and (b) incidents of household property crime per 1,000 households interviewed each month



Source: United States Census Bureau, *National Crime Victimization Survey*, 2008–2012

The first survey outcome is the rate of personal crime incidents. This rate is defined as the total number of crimes committed against persons (including personal theft and violent crime) collected by NCVS interviewers during a specified period of time, divided by the number of persons interviewed during that period of time. A plot of these monthly rates of personal crime is shown in Figure 5 (a). These rates are smaller than the victimization rates published annually by BJS. The published rates refer to a calendar year, whereas the rates shown in Figure 5 (a) refer to the six-month window prior to the date of the interview. Published victimization rates are weighted to take into account sample selection procedures, nonresponse, differing windows of time, and the fact that some crimes have multiple victims. Nevertheless, the two measures are closely related, and changes in the rates shown in Figure 5 (a) will strongly affect the published rates. The rates shown in Figure 5 (a) are generally rising from 2010 to 2012, with fluctuations due to possible seasonal effects and noise.

Our second survey outcome is the rate of household property crime incidents. This variable, plotted in Figure 5 (b), is the number of household property crimes collected by interviewers divided by the number of households interviewed. This rate also appears to rise from 2010 to 2012, with possible seasonal effects and noise.

Our previous analyses showed that refresher training and enhanced monitoring increased the apparent rates of personal crime and household property crime during the latter months of 2011, but only among those crimes that had not been reported to police [1]. In the sections ahead, we will again distinguish crimes by whether or not they were reported to police. For certain categories of crime, unreported crimes tend to be less serious and harder for respondents to recall. Crimes reported to police may be more salient in respondents' memories and more likely to be reported to FRs during the screener interviews, whereas unreported crimes may be less salient and more susceptible to interviewer effects and changes in the field conditions. Relationships between salience and difficulty of recall have been demonstrated by Miller and Groves (1985) [2] and by Czaja et al. (1994) [3].

3 MODELING STRATEGY

Basic Form of the Models

The variables plotted in Figures 4 and 5 are summary measures for each month. But the interventions of interest, refresher training/performance monitoring and field realignment, took effect in different months for different FRs. This variation in timing across FRs is a key part of our strategy for estimating intervention effects. In a sense, this variation in timing across FRs provides the replication that we need to separate the effects of the intervention from long-term trends and seasonal shifts that may be happening at the same time. To tease out the effects of interventions from other temporal phenomena, we must disaggregate the data by interviewers. The basic unit of analysis for each of our models is the interviewer-month.

Measurements for interviewer-months are severely unbalanced. From 2008 to 2012, the number of interviewers working for the NCVS in any given month was approximately 650–800. Over the entire period, however, about 1,900 different interviewers worked on the survey. Some highly experienced FRs were present for all five years, but many worked on the survey only sporadically, and some were present for only a single month. Methods that require a complete or nearly complete series for each interviewer will not be appropriate for these data. Fortunately, the models that we fit are not adversely affected by the lack of balance. Each interviewer contributes observations for however many months he or she worked for the NCVS from 2008 to 2012, and where appropriate, each month's contribution is weighted by the interviewer's NCVS caseload during that month.

Each of our models will take the following general form. Let Y_{ij} denote an aggregated outcome (e.g., response rate) for interviewer i during month t_j. We assume that

$$Y_{ij} \sim F(\mu_{ij};\, \phi, w_{ij}), \qquad (1)$$

where F is a user-specified parametric family with mean $\mu_{ij} = E(Y_{ij})$, optional dispersion parameter ϕ, and optional inverse-variance weight w_{ij}. The mean is decomposed as

$$g(\mu_{ij}) = \omega_{ij} + f_1(t_j) + f_2(t_j) + \alpha_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}, \qquad (2)$$

where

- g is a user-specified link function,
- ω_{ij} is an optional offset term needed for rate models,
- f_1 is a smooth function describing a long-term trend,
- f_2 is a periodic function describing an annual cycle,
- α_i is a random effect for interviewer i, assumed to be distributed as $N(0, \sigma_\alpha^2)$,
- $\mathbf{x}_{ij}$ is a vector of covariates, and
- $\boldsymbol{\beta}$ is a vector of coefficients to be estimated.

The effects of field interventions are contained in $\boldsymbol{\beta}$, and their interpretation will depend on how the variables in $\mathbf{x}_{ij}$ are coded. The reliability of those estimates will depend on how accurately we describe the long-term and annual trends in f_1 and f_2. For our purposes, it is best to avoid simple parametric forms (e.g., linear or quadratic functions of time) which are probably unrealistic, and work with function classes that are more general. Before listing the covariates in $\mathbf{x}_{ij}$, we describe our strategy for characterizing f_1 and f_2.

Approximating a Long-Term Trend Using a Natural Cubic Spline

Splines are classes of functions that are flexible enough to approximate a trend of almost any shape. To introduce the idea, suppose that we have a data series $y_1, \dots, y_n$ that represents a response variable recorded at times $t_1, \dots, t_n$. Suppose that

$$y_i = f(t_i) + \epsilon_i,$$

where the ϵ_i's are independent errors with mean zero and constant variance. Our goal is to estimate f, a function that is believed to be continuous and smooth but whose shape is not otherwise known. To create a spline, we first partition the real line into $K+1$ intervals,

$$(-\infty, \xi_1), [\xi_1, \xi_2), \dots, [\xi_K, \infty).$$

where the cutpoints $\xi_1 < \xi_2 < \cdots < \xi_K$ are called knots. A spline of degree r consists of an rth degree polynomial over each interval,

$$f(t) = \begin{cases} \beta_{00} + \beta_{01}\, t + \dots + \beta_{0r}\, t^r & \text{if } t \in (-\infty, \xi_1), \\ \beta_{10} + \beta_{11}\, t + \dots + \beta_{1r}\, t^r & \text{if } t \in [\xi_1, \xi_2), \\ \vdots & \\ \beta_{K0} + \beta_{K1}\, t + \dots + \beta_{Kr}\, t^r & \text{if } t \in [\xi_K, \infty). \end{cases}$$

To make the function continuous and smooth, we will apply constraints to the β's to force f and its first $r-1$ derivatives to be continuous at the knots. Various methods for constructing splines are available. One simple method relies on the truncated power basis. In this method, a spline of degree r with knots at $\xi_1, \dots, \xi_K$ is written as

$$\begin{aligned} f(t) &= \beta_0 + \beta_1\, t + \cdots + \beta_r\, t^r \\ &\quad + u_1\, (t-\xi_1)_+^r + \cdots + u_K\, (t-\xi_K)_+^r, \end{aligned}$$

where

$$(t-\xi_j)_+^r = \begin{cases} (t-\xi_j)^r & \text{if } t \geq \xi_j, \\ 0 & \text{otherwise.} \end{cases}$$

To estimate f, we compute the $r+K+1$ basis functions

$$1,\ t,\ \dots,\ t^r,\ (t-\xi_1)_+^r, \dots, (t-\xi_K)_+^r$$

for $t_1, \dots, t_n$ and treat them as regressors, fitting an ordinary least-squares (OLS) regression of $y_1, \dots, y_n$ on these variables to estimate the β's and u's. If a linear regression is not appropriate (e.g., if the y_i represents a proportion or rate) then we can easily switch to a generalized linear model (e.g., logistic or loglinear regression).

To illustrate, Figure 6 shows an artificial dataset of $n = 26$ observations whose measurement times are equally spaced. The line plotted in Figure 6 (a) is a cubic spline with $K = 5$ equally spaced knots fit by OLS. Cubic splines are a popular choice, because they are very smooth; the knots in the fitted curve are invisible to the eye. One disadvantage shown by this example is that at the ends of the series, the fitted function becomes erratic; beyond the boundary knots (where $t < \xi_i$ and $t > \xi_k$), the estimate of $f(t)$ is strongly pulled toward the first and last observations (y_1 and y_n), causing the curve to veer off in implausible directions.

To stabilize a cubic spline near the edges of the space, it is customary to impose additional constraints to require the function to be linear beyond the boundary knots,

$$f''(t) = 0 \quad \text{for } t \leq \xi_1 \text{ and } t \geq \xi_K.$$

A cubic spline with this constraint is called a natural cubic spline. Natural cubic splines can be constructed in various ways. For our purposes, we will use second divided differences of truncated power functions (White *et al.*, 1998; Welham, 2008) [4] [5]. Given a set of knots $\xi_1, \dots, \xi_k$, the natural cubic spline is written as

$$f(t) = \beta_0 + \beta_1\, t + \sum_{k=2}^{K-1} u_k\, P_k^*(t),$$

where

$$\begin{aligned} P_k^*(t) &= \frac{1}{6} \left\{ h_k^{-1}\, (t-\xi_{k+1})_+^3 - (h_k^{-1} + h_{k-1}^{-1})\, (t-\xi_k)_+^3 \right. \\ &\qquad \left. + h_{k-1}^{-1}\, (t-\xi_{k-1})_+^3 \right\}, \end{aligned}$$

and $h_k = (\xi_{k+1} - \xi_k)$. To reduce collinearity, we replace $P_k^*(t)$ with

$$P_k(t) = P_k^*(t) - \hat{a}_k - \hat{b}_k\, t,$$

where $\hat{a}_k$ and $\hat{b}_k$ are the intercept and slope from the ordinary least-squares regression of $P_k^*(t)$ on t over the sample points,

$$\begin{bmatrix} \hat{a}_k \\ \hat{b}_k \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n P_k^*(t_i) \\ \sum_{i=1}^n t_i\, P_k^*(t_i) \end{bmatrix}.$$

The basis becomes

$$f(t) = \beta_0 + \beta_1\, t + \sum_{k=2}^{K-1} u_k\, P_k(t). \qquad (3)$$

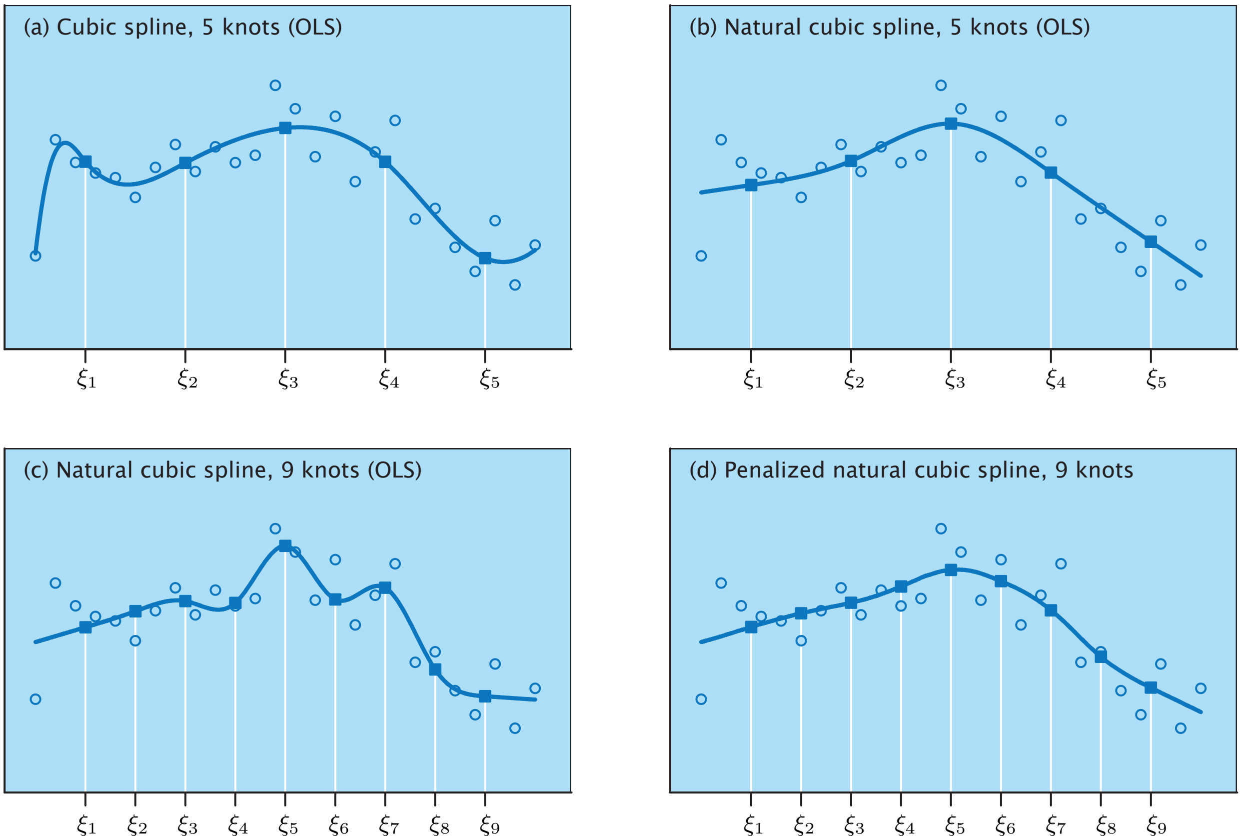
This representation has the attractive property that

$$u_k = f''(\xi_k)$$

Figure 6

Examples of Fitted Spline Functions

Hypothetical data series of $n = 26$ equally-spaced observations, with estimated mean functions approximated by four different methods: (a) cubic spline with $K = 5$ knots, fit by ordinary least squares; (b) natural cubic spline with $K = 5$ knots, fit by ordinary least squares; (c) natural cubic spline with $K = 9$ knots, fit by ordinary least squares; and (d) penalized natural cubic spline with $K = 9$ knots, fit as a linear mixed model



for each interior knot $k = 2, \dots, K - 1$. The second derivatives vanish at the boundary knots,

$$f''(\xi_1) = f''(\xi_K) = 0,$$

in keeping with the definition of the natural cubic spline. A natural cubic spline with $K = 5$ knots fit to the artificial dataset by OLS is shown in Figure 6 (b). The extra constraints of linearity beyond the boundary knots have solved the problem of instability at the edges.

Roughness Penalties and Mixed Models

The smoothness of a spline is heavily influenced by the number and placement of knots. As the number of knots increases, the curve becomes more flexible, exhibiting short-term peaks and valleys. A natural cubic spline with $K = 9$ equally spaced knots fit by OLS is shown in Figure 6 (c). Comparing Figures 6 (b) and 6 (c), the latter shows more fluctuation. The spline with $K = 9$ knots appears to be overfitted; the short-term fluctuations are probably just noise. However, it is possible that the spline with $K = 5$ knots is underfitted, glossing over some features that may be authentic.

For scientific reasons, we want to avoid subjective choices about the number and placement of knots that will heavily influence the results. Rather, we prefer a method that adapts automatically, choosing a degree of smoothness appropriate for the given dataset. This problem has been well studied, with many articles and books published over several decades. One well known approach is the smoothing spline (Wahba, 1990; Green and Silverman, 1994) [6] [7]. The smoothing spline is defined as the minimizer of the function

$$(\mathbf{y} - \mathbf{f})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{f}) + \lambda \int [f''(t)]^2 \, dt \qquad (4)$$

over the space of twice-differentiable functions f, where $\mathbf{y} = (y_1, \dots, y_n)^T$, $\mathbf{f} = (f(t_1), \dots, f(t_n))^T$, $\mathbf{V}$ is the covariance matrix of $Y_1, \dots, Y_n$, and λ is a user-specified smoothing parameter. The solution to this minimization problem is a natural cubic spline with knots located at the design points (i.e., at all the distinct values of t_i).

Smoothing splines have attractive theoretical properties, but computing them becomes time consuming as the number of design points increases. An economical alternative is to specify a grid of knots, spaced equally over the range of $t_1, \dots, t_n$ or at their quantiles, and fit a spline with a roughness penalty. Variations on this approach are known as P-splines (Eilers and Marx, 1996) [8] and penalized splines (Ruppert, Wand and Carroll, 2003) [9].

Consider a natural cubic spline model

$$y_i \;=\; \beta_0 \;+\; \beta_1\, t_i \;+\; \sum_{k=2}^{K-1} u_k\, P_k(t_i) \;+\; \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, and where the P_k's are the basis functions shown in (3). Suppose we apply a large number of knots $\xi_1, \dots, \xi_K$ spaced at regular intervals. The number of knots is not crucial, provided that it is large enough that the ordinary regression estimate will be overfitted. Estimating the β's and u's by OLS will produce a fitted curve that exhibits too much fluctuation. However, if we treat the u's as random variables drawn from a common distribution,

$$u_2, \dots, u_{K-1} \;\sim\; N(0, \sigma_u^2), \tag{5}$$

then we can estimate the additional variance component σ_u^2 from the data. The model becomes

$$\mathbf{y} \;=\; \mathbf{X}\boldsymbol{\beta} \;+\; \mathbf{P}\mathbf{u} \;+\; \boldsymbol{\epsilon}, \tag{6}$$

where

$$\mathbf{X} \;=\; \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}, \quad \mathbf{P} \;=\; \begin{bmatrix} P_2(t_1) & \dots & P_{K-1}(t_1) \\ P_2(t_2) & \dots & P_{K-1}(t_2) \\ \vdots & \ddots & \vdots \\ P_2(t_n) & \dots & P_{K-1}(t_n) \end{bmatrix},$$

$\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, $\mathbf{u} = (u_2, u_3, \dots, u_{k-1})^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$, and where

$$\begin{aligned} \mathbf{u} \;&\sim\; N(0, \sigma_u^2 \mathbf{I}), \\ \boldsymbol{\epsilon} \;&\sim\; N(0, \sigma_\epsilon^2 \mathbf{V}). \end{aligned}$$

This is an example of a linear mixed model, also known as a linear mixed-effects model. It differs from common examples of linear mixed models in that the observational units $j = 1, \dots, n$ are crossed with the random coefficients rather than nested within them. Nevertheless, this model can be fit with software packages that accommodate crossed random effects (e.g., PROC MIXED in SAS). Estimates of the random coefficients in $\mathbf{u}$ will be shrunk toward zero, producing a curve that is smoother than the OLS version. In fact, under these basis functions, the u_j's represent the second derivatives of f at the interior knots, and the distributional assumption on $\mathbf{u}$ imposes a roughness penalty that is a discrete approximation to the second term in the smoothing spline (4). The fitted curve from this penalized natural cubic spline is a data-determined compromise between the overfitted OLS spline and a simple linear regression of $y_1, \dots, y_n$ on $t_1, \dots, t_n$.

To illustrate, we fit a penalized natural cubic spline with $K = 9$ knots to the artificial dataset shown in Figure 6, using the linear mixed-model formulation (6). The resulting curve is plotted in Figure 6 (d). The extra distributional assumption (5) imposed on the u's has effectively removed the short-term peaks and valleys found in the OLS curve of Figure 6 (c).

In the following way, we embed penalized natural cubic splines into our models for NCVS data. Returning to the model shown in Equation (2), we represent the long-term trend as

$$f_1(t) \;=\; \beta_0 \;+\; \beta_1\, t \;+\; \sum_{k=2}^{K-1} u_k\, P_k(t), \tag{7}$$

with knots $\xi_1, \dots, \xi_K$ placed at the beginning of each quarter-year. The first two regressors (the constant and t) are placed into the covariate vector $\mathbf{x}_{ij}$, so that β_0 and β_1 are subsumed into $\boldsymbol{\beta}$. The basis functions $P_2(t), \dots, P_{K-1}(t)$ are added to the model (2) as regressors with random coefficients distributed as $u_2, \dots, u_{K-1} \sim N(0, \sigma_u^2)$. The resulting model becomes a generalized linear mixed model with observational units that are crossed with the random coefficients $u_2, \dots, u_{K-1}$.

Approximating a periodic trend by a periodic cubic spline

In addition to the long-term trend f_1, we need a method for specifying the function f_2 in our model (2) that is both flexible and periodic, with a period fixed at one year. That is, if t is expressed in years, we need $f_2(t)$ to be continuous and smooth, with the property

$$f_2(t) \;=\; f_2(t+1) \;=\; f_2(t+2) \;= \cdots$$

for any value of t.

Various methods for incorporating periodic trends into longitudinal models have been proposed. One popular technique is to apply a Fourier basis consisting of sine-cosine pairs. Using a Fourier basis, a periodic function with period T can be expressed as

$$f_2(t) \;=\; \sum_{m=1}^{M} \left\{ \delta_{2m-1} \sin\left(\frac{2\pi mt}{T}\right) \,+\, \delta_{2m} \cos\left(\frac{2\pi mt}{T}\right) \right\},$$

where $\delta_1, \ldots, \delta_{2M}$ are coefficients to be estimated. With monthly measurements, we have $12 - 1 = 11$ degrees of freedom available to estimate an annual cycle, so the largest number of sine-cosine pairs we can use is $M = 5$. While experimenting with Fourier bases, we found that the resulting fitted function f_2 exhibited implausible short-term oscillations within the year. To dampen these short-term oscillations, we switched to a method based on periodic cubic splines (Zhang, Lin and Sowers, 2000) [10].

To construct a periodic spline with period T, we begin with the ordinary cubic spline

$$f_2(t) \;=\; \beta_0 \,+\, \beta_1\, t \,+\, \beta_2\, t^2 \,+\, \beta_3\, t^3 \,+\, \sum_{k=1}^{K} \delta_j \,(t-\xi_j)_+^3 \quad (8)$$

for coefficients $\beta_0, \ldots, \beta_3, \delta_1, \ldots, \delta_K$, with knots located between 0 and T. To make the spline periodic over $t \in [0, \infty)$, two changes are needed. First, we replace each occurrence of t on the right-hand side of (8) with the seasonal operator

$$s(t) \;=\; \mathrm{mod}(t, T).$$

Second, we enforce continuity upon f and its first two derivatives by imposing the constraints $f_2(0) = f_2(T)$, $f_2'(0) = f_2'(T)$, and $f_2''(0) = f''(T)$. With some algebra, the constrained version of (8) becomes

$$f_2(t) \;=\; \beta_0 \,+\, \sum_{k=1}^{K} \delta_k \, P_k^*(\, s(t)\,),$$

where

$$P_k^*(s) \;=\; a_k s + b_k s^2 + c_k s^3 + (s - \xi_k)_+^3,$$

and

$$\begin{aligned} a_k &= -\frac{T(T-\xi_k)}{2} + \frac{3(T-t_k)^2}{2} - \frac{(T-\xi_k)^3}{T}, \\ b_k &= \frac{3(T-\xi_k)}{2} - \frac{3(T-\xi_k)^2}{2\,T}, \\ c_k &= -\frac{T-\xi_k}{T}, \end{aligned}$$

which corrects a typographical error by Zhang, Lin and Sowers (2000) [10]. Finally, we impose an additional requirement that

$$\int f_2(t)\, dt \;\approx\; 0$$

by setting $\beta_0 = 0$ and centering each basis function at its average value, averaging over the design points of t. Let $t_1, t_2, \ldots, t_n$ denote the distinct values of t appearing in the dataset, and let

$$P_k(s) \;=\; P_k^*(s) \;-\; \bar{P}_k,$$

where

$$\bar{P}_k \;=\; \frac{1}{n} \sum_{i=1}^{n} P_k^*(\, s(t_i)\,).$$

The function becomes

$$f_2(t) \;=\; \sum_{k=1}^{K} \xi_k \, P_k(\, s(t)\,).$$

We found that by placing $K = 6$ knots over the calendar year at two-month intervals, we have been able capture the major features of each periodic trend, but without the implausible within-year oscillations that were an artifact of the Fourier basis. These six periodic basis functions are placed in the covariate vector $\mathbf{x}_{ij}$ of our model (2), and the corresponding coefficients are subsumed into β.

Intervention Effects

Having described the long-term and periodic trends, we now turn our attention to the effects of the interventions. After experimenting with different ways to code the interventions, we settled on a very simple method. To account for refresher training and performance monitoring, we created a pair of dummy indicators,

$$\begin{aligned} \text{After.Training.2011}_{ij} &= \text{1 if } t_j \text{ is in 2011 and interviewer } i \text{ had been trained, 0 otherwise,} \\ \text{After.Training.2012}_{ij} &= \text{1 if } t_j \text{ is in 2012 and interviewer } i \text{ had been trained, 0 otherwise,} \end{aligned}$$

and included them as covariates in the vector $\mathbf{x}_{ij}$. The corresponding elements of β become simple contrasts (trained minus untrained) for 2011 and 2012.

Similarly, we defined a dummy indicator for field realignment,

$$\text{After.Realignment}_{ij} \;=\; \text{1 if interviewer } i \text{ experienced realignment by month } t_j\text{, 0 otherwise,}$$

so that the corresponding element of β is a simple contrast (realigned versus not).

Information for estimating these coefficients comes mainly from data in late 2011 and 2012 when the interventions are being phased in. Data from earlier years may strengthen these estimates by providing information about long-term and periodic trends and the effects of other covariates.

Additional Covariates

In addition to the long-term and periodic trends, our models include additional covariates to adjust for differences in the interviewers' monthly assignments. One key covariate, which we described in Section 2, is the interviewer's caseload,

$$\text{WLHH}_{ij} \quad = \quad \text{number of NCVS HH's assigned to interviewer } i \text{ during month } t_j.$$

In each of our models, this variable is included as a covariate in $\mathbf{x}_{ij}$. Depending on the model, it may also appear as a precision weight w_{ij} or as an offset term ω_{ij}.

The other covariate in our models comes from the Census Bureau's Planning Database. The Planning Database, which was created using data from the 2010 Census and the American Community Survey, provides information at fine levels of geography (census block groups) that are known to be helpful for planning and designing surveys. One such variable is a national classification of block-groups into three clusters that represent different levels of difficulty for conducting census and survey operations. Cluster 1 represents areas where enumeration is easy and response rates tend to be high. Rural and suburban locations tend to fall into Cluster 1. Clusters 2 and 3 represent more heavily urbanized areas where data collection tends to be more difficult. By linking to the Planning Database, we were able to identify the cluster membership for about 88% of the housing units in NCVS. For the remaining 12%, the cluster could not be determined because key geographic information was missing. Exploratory analyses showed that among those 12%, the NCVS response rates were unusually low, so we decided to combine them with Clusters 2 and 3. We then aggregated this variable to produce a summary statistic for the interviewer-month,

$$\text{Cluster.1}_{ij} \quad = \quad \text{proportion of interviewer } i\text{'s cases during month } t_j \text{ that are known to belong to Planning Database Cluster 1.}$$

We conjectured that this covariate would have a significant positive correlation with response rates and a significant negative correlation to crime rates.

Fitting the Models

The models we have described are generalized linear mixed models with an unusual pattern of crossed and nested random effects. Although some existing software packages (PROC NLMIXED in SAS, the GLAMM package in Stata) are capable of fitting models like these, we could not find any available package that had all the capabilities we needed. With observations for approximately 42,000 interviewer-months, it is difficult to fit these models in a reasonable amount of time.

To compute parameter estimates, we implemented a specialized Bayesian Markov chain Monte Carlo (MCMC) procedure programmed in Fortran 95 and called from R. For each model, the MCMC chain was run for 25,000 cycles following a burn-in period of 1,000 cycles. Parameters were subsampled and saved with a thinning interval of 5 cycles, yielding a sample of 5,000 draws from the posterior distribution.

For an overview of the algorithm and the prior distributions applied to the parameters, see the Technical Appendix by Schafer (2013) [1].

4 RESULTS FOR HOUSEHOLD RESPONSE RATES

In our first model, the outcome variable is

$$Y_{ij} \quad = \quad \text{household response rate for interviewer } i \text{ during month } t_j.$$

We assumed a binomial distribution,

$$Y_{ij} \sim n_{ij}^{-1}\,\text{Bin}(\,n_{ij},\,\mu_{ij}),$$

where $n_{ij} = \text{WLHH}_{ij}$ is the number of households assigned to interviewer i during month t_j. The link function is logistic, and the mean function is

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) \quad = \quad f_1(t_j) + f_2(t_j) + \alpha_i + \mathbf{x}_{ij}^T\boldsymbol{\beta},$$

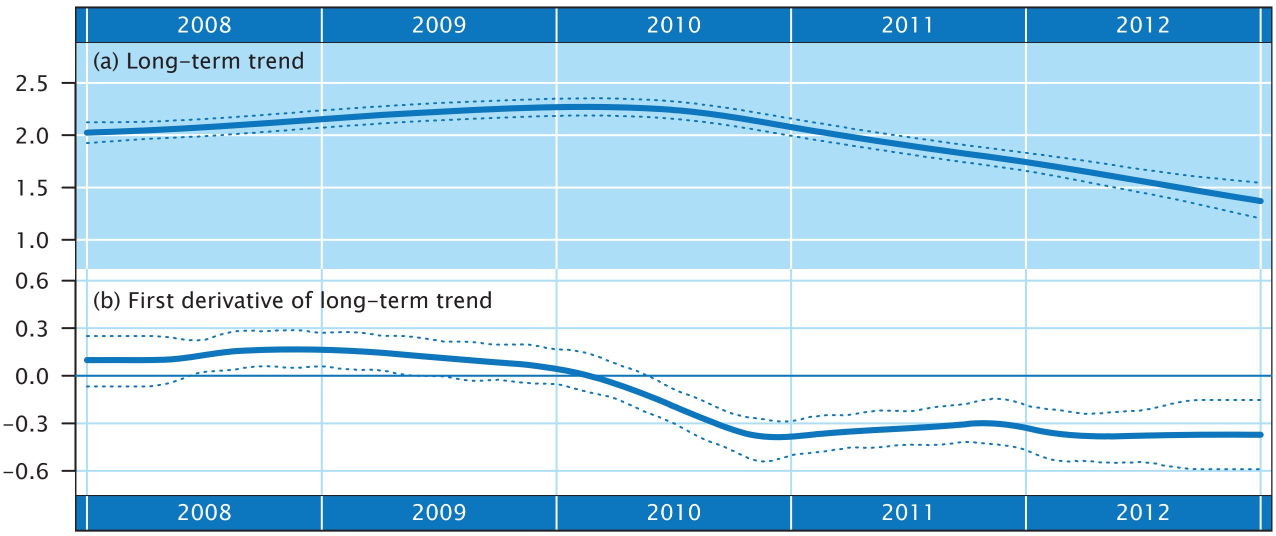
where f_1 is a long-term trend, f_2 is an annual periodic cycle, α_i is a random effect for interviewer i, $\mathbf{x}_{ij}$ is a vector of covariates, and β is a vector of coefficients.

The estimate of the long-term trend is plotted in Figure 7 (a), along with pointwise 95% error

Figure 7

Long-Term Trend for Household Response Rate

(a) Estimated long-term trend for household response rate on the log-odds scale (solid line) with pointwise 95% error bounds (dashed lines), and (b) estimated first derivative of the long-term trend (solid line) with pointwise 95% error bounds (dashed lines)



Source: United States Census Bureau, *National Crime Victimization Survey*, 2008–2012

bounds. At each value of t, the estimate value of $f_1(t)$ is a simulated posterior median, and the error bounds are the 2.5th and 97.5th posterior percentiles, based on sample of 5,000 draws from the joint posterior distribution of the unknown coefficients in (7). This trend, which is displayed on the logistic or log-odds scale, tracks the overall pattern seen in Figure 4 (a); the response rate rose from 2008 through the first quarter of 2010 and has been declining ever since. Notice that the first derivative of the spline function (7) is

$$f_1'(t) \;=\; \beta_1 \;+\; \sum_{k=2}^{K-1} u_k\, P_k'(t), \tag{9}$$

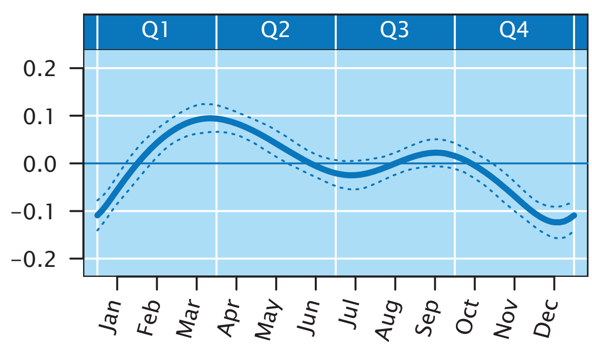
where $P_k'(t)$ is the first derivative of the natural cubic spline basis function $P_k(t)$ defined in Section 3. Under this model, estimating the instantaneous change is no more difficult than estimating the trend itself, because both the natural cubic spline and its derivative are linear combinations of the same coefficients. The estimated first derivative is plotted in Figure 7 (b), along with 95% pointwise error bounds. From early 2010 onward, the estimate of $f_1'(t)$ has been consistently negative, and the error bounds lie below zero, indicating that throughout this period the downward trend in response rate was statistically significant.

A plot of the estimated annual periodic cycle $f_2(t)$ is shown in Figure 8, along with 95% error bounds. Because the average value of this function over the year is constrained to be zero, error bounds that do not cover zero indicate times where the response rate lies above or below its annual average. The major features of this periodic cycle are a significant increase in response rate during March and April and a significant decrease during the holiday season at the end of the year.

Figure 8

Annual Periodic Cycle for Household Response Rate

Estimated annual periodic cycle for household response rate on the log-odds scale (solid line) with pointwise 95% error bounds (dashed lines)



Source: United States Census Bureau, *National Crime Victimization Survey*, 2011–2012

Table 1: Coefficients, standard errors, and Bayesian p-values from model for household response rate

	Coef	SE	p
WLHH	.0011	.0009	.224
Cluster.1	.0682	.0240	.004
After.Training.2011	−.0847	.0335	.014
After.Training.2012	−.0677	.0390	.078
After.Realignment	.0760	.0459	.096

Source: United State Census Bureau, *National Crime Victimization Survey*, 2008–2012

Estimates for the key coefficients in β are shown in Table 1, along with standard errors and simulated p-values. In this table and all subsequent tables, the estimated coeffients and standard errors are simulated posterior means and standard deviations, averaged across 5,000 random draws from the Bayesian posterior distribution. The p-values are simulated Bayesian p-values, defined as one minus the probability content of the narrowest equal-tailed posterior interval that covers the parameter's null value of zero. These may be interpreted in roughly the same manner as significance values from frequentist two-tailed hypothesis tests, with a value of .05 or less indicating an effect that is statistically significant.

In Table 1, the coefficient for WLHH (.0011) is not significantly different from zero ($p = .224$), indicating that there is little evidence of a relationship between response rates and interviewers' monthly workloads.

The coefficient for Cluster.1 (.0682) is positive and statistically significant ($p = .004$). As expected, a higher proportion of households located in Planning Database Cluster 1 is associated with a higher response rate.

Examining the coefficient for After.Training.2011 (−.0847), we see that the estimated effect of refresher training and performance monitoring in 2011 is negative and statistically significant ($p = .014$). The corresponding effect for 2012 ((−.0677)) is not significant ($p = .078$). These coefficients pertain to the log-odds. To understand the implications on the probability scale, suppose we start with a response rate of 90%. A change in the log-odds of −0.0847 would reduce the rate by 0.8 percentage points to 89.2%, and a change in the log-odds of −0.0677 reduces the rate by 0.6 percentage points to 89.4%. Thus, refresher training and performance monitoring are associated with a modest but significant decrease in household response rates in 2011, and a smaller, non-significant decrease in 2012.

One possible explanation for why refresher training and performance monitoring would reduce response rates is the change in perfomance standards that were discussed in Section 2. Prior to training, FRs were evaluated solely on their response rates. As the interviewers were trained, managers were instructed to begin monitoring their performance by additional quality measures that included screener times. If FRs had previously been trying to convert difficult cases (households for which response was unlikely) by conducting the screeners too quickly, then we would expect response rates under the new performance measures to go down. The fact that they have gone down only slightly is relatively good news. It seems plausible that the new measures had the intended consequence of emphasizing survey quality over just response rates, possibly reducing falsified and quick interviews.

The estimated coefficient for After.Realignment (.0760) is positive. Field realignment was implemented during 2012, a period when response rates were declining. If this effect were real, it would indicate that realignment slowed that decline, and the response rate after realignment was slightly higher (by approximately 0.7 percentage points) than it would have been without realignment. However, this effect is not statistically significant ($p = .096$), so the evidence for an effect of realignment is inconclusive.

5 RESULTS FOR SCREENER TIMES

For our second model, the response variable is

$$Y_{ij} \;=\; \text{average screener time for interviewer } i \text{ during month } t_j.$$

We assumed a normal distribution,

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2/n_{ij}),$$

where n_{ij} is the number of persons interviewed by interviewer i during month t_j. The link function is the identity. Preliminary analyses revealed that there were no discernible annual periodic effects in screener times, so we simplified this model by removing the periodic component. The model is

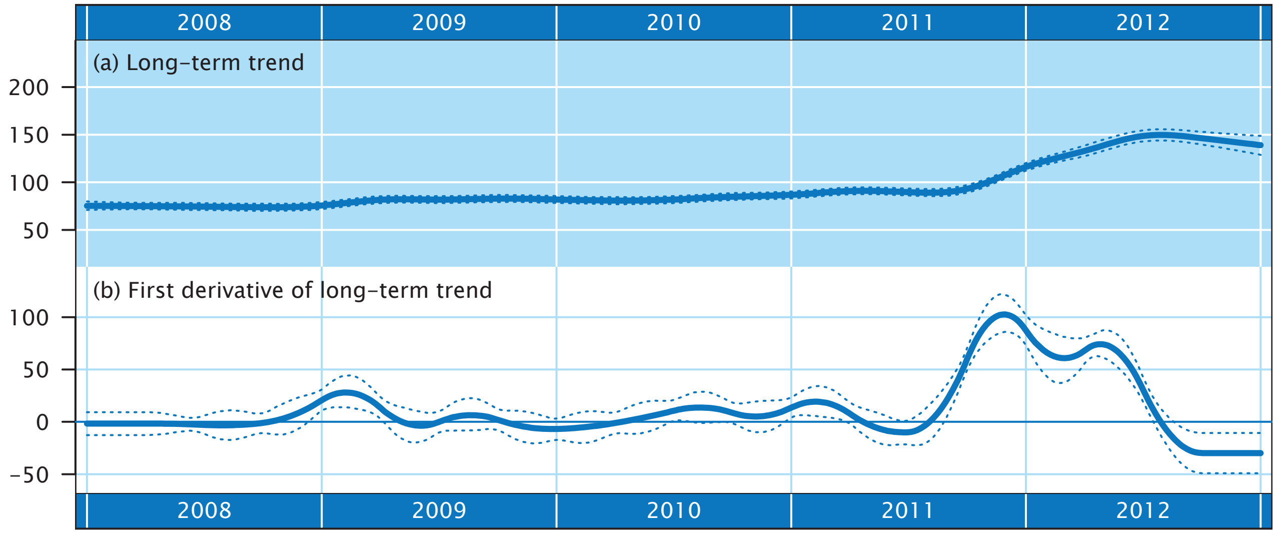
$$\mu_{ij} \;=\; f_1(t_j) + \alpha_i + \mathbf{x}_{ij}^T \boldsymbol{\beta},$$

where f_1 is a long-term trend, α_i is a random effect for interviewer i, $\mathbf{x}_{ij}$ is a vector of covariates, and β is a vector of coefficients.

Figure 9

Long-Term Trend for Average Screener Time

(a) Estimated long-term trend for average duration of the screener interview in seconds (solid line) with pointwise 95% error bounds (dashed lines), and (b) estimated first derivative of the long-term trend (solid line) with pointwise 95% error bounds (dashed lines)



Source: United States Census Bureau, *National Crime Victimization Survey*, 2008–2012

The estimated long-term trend $f_1(t)$ is shown in Figure 9. This plot shows a large increase in screener times of about 60–70 seconds in late 2011 and early 2012, the period when refresher training and performance monitoring was being phased in. This effect is dramatic, but it is smaller than the increase in screener times of approximately two minutes that we saw in Figure 4 (b). To understand the reason for the discrepancy, note that this model also includes as predictors the dummy-indicator variables After.Training.2011 and After.Training.2012. The coefficients for those predictors, which will be shown momentarily, estimate the jump in screener times for an FR immediately after the FR was trained. In addition to that immediate increase, the long-term trend suggests that all FRs experienced a general pattern of increase in late 2011 and early 2012 regardless of when they were trained. This might be explained by the new performance monitoring measures which were introduced across the workforce, including a new benchmark of 3.5 minutes (210 seconds) for screener interviews. Because training and enhanced monitoring were enacted at roughly the same time, it is difficult to separate the effects of these two initiatives. The longitudinal model appears to have picked up this combined effect partly in the coefficients for After.Training.2011 and After.Training.2012, and partly in the estimated long-term trend.

The first derivative of the long-term trend with 95% error bounds is shown in Figure 9 (b). This plot reveals brief periods of significant increase in early 2009, mid-2010 and early 2011; a period of major increase in late 2011 and early 2012; and a period of significant decline in late 2012. The declining screener times at the end of the series may be a cause for concern, depending on what happens in 2013.

Table 2 shows coefficients, standard errors and p-values from the screener-times model. The coefficient for WLHH is negative and statistically significant ($p = .000$). Interviewers with higher monthly workloads tend to spend less time on screener interviews than those with lower monthly workloads. The size of this effect, however, is very small. The coefficient of $-.1838$ suggests that, holding all other predictors constant, an increase in workload of one household per month would reduce the average screener time by only about one-fifth of a second, a change that has no practical significance.

In contrast, the estimated effects for After.Training.2011 and After.Training.2012 (69.28 and 74.05) are both statistically significant and large. As mentioned earlier, these estimate the change in average screener time that occurred immediately after training. They suggest that, if all other explanatory variables were equal, the estimated difference in average screener times for trained and untrained interviewers would be about

Table 2: Coefficients, standard errors, and Bayesian p-values from model for average screener time

	Coef	SE	p
WLHH	−.1838	.0390	.000
Cluster.1	−.9582	1.137	.392
After.Training.2011	69.28	1.798	.000
After.Training.2012	74.05	2.384	.000
After.Realignment	3.538	2.849	.214

Source: United State Census Bureau, *National Crime Victimization Survey*, 2008–2012

70 seconds. Over the same period of time when training was taking place, however, the long-term trend showed an average increase of roughly 60-70 seconds. Taken together, these effects account for an increase in average screener times by about two minutes that were seen in late 2011 and early 2012 as refresher training and enhanced monitoring were being phased in.

6 RESULTS FOR PERSONAL CRIME

Our next model describes the number of incidents of personal crime recorded during the interview process. The response variable is

$$Y_{ij} \;=\; \text{personal crimes discovered by interviewer } i \text{ during month } t_j.$$

This variable is assumed to have a negative binomial distribution,

$$Y_{ij} \sim \text{NegBin}(\,\alpha = \kappa^{-1},\, \beta = \kappa^{-1}\mu_{ij}\,),$$

where $\kappa > 0$ is an unknown dispersion parameter. We applied a negative binomial model because preliminary analyses showed that the responses were overdispersed relative to a Poisson distribution.

It is reasonable to believe that the number of personal crimes reported by an interviewer is approximately proportional to the number of persons interviewed. That is, we may suppose

$$\mu_{ij} \propto n_{ij},$$

where n_{ij} is the number of persons interviewed by interviewer i during month t_j. Applying this assumption, and using a logarithmic link, the model for the mean becomes

$$\log \mu_{ij} \;=\; \omega_{ij} + f_1(t_j) + f_2(t_j) + \alpha_i + \mathbf{x}_{ij}^T \boldsymbol{\beta},$$

where $\omega_{ij} = \log n_{ij}$, $f_1(t)$ and $f_2(t)$ are long-term and periodic trends, α_i is a random effect for interviewer i, and $\mathbf{x}_{ij}$ is a vector of covariates. The ω_{ij} on the right-hand side of the equation is called an offset term; it is a predictor whose coefficient is assumed to be fixed at one. Alternatively, we can place the offset to the left-hand side, so that this can be viewed as a loglinear model for the personal crime incident rate,

$$\log\left(\frac{\mu_{ij}}{n_{ij}}\right) \;=\; f_1(t_j) + f_2(t_j) + \alpha_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}. \quad (10)$$

Plots of the estimated long-term trend $f_t(t)$, and the estimated first derivative of the estimated long-term trend $f_1'(t)$, are shown in Figure 10. The estimate of $f_1(t)$ shows a mild decrease from 2008 to 2010 and a mild increase from 2010 to 2012. Except for a period in 2009 where the error bounds on $f_1'(t)$ briefly dip below zero, the evidence for change is not conclusive. However, the overall pattern is consistent with the official victimization estimates reported annually by BJS. Except for small numbers of personal thefts (pocket picking, completed or attempted purse snatching), most of the crime incidents included in this model were classified as violent crimes. Truman, Langton and Planty (2013) reported that rates of violent crime victimization rose over the last two years (from 2010 to 2011, and from 2011 to 2012) after a period of steady decline prior to 2010 [11]. The curve shown in Figure 10 (a) has the same shape: declining from 2008 to 2010, rising from 2010 to 2012.

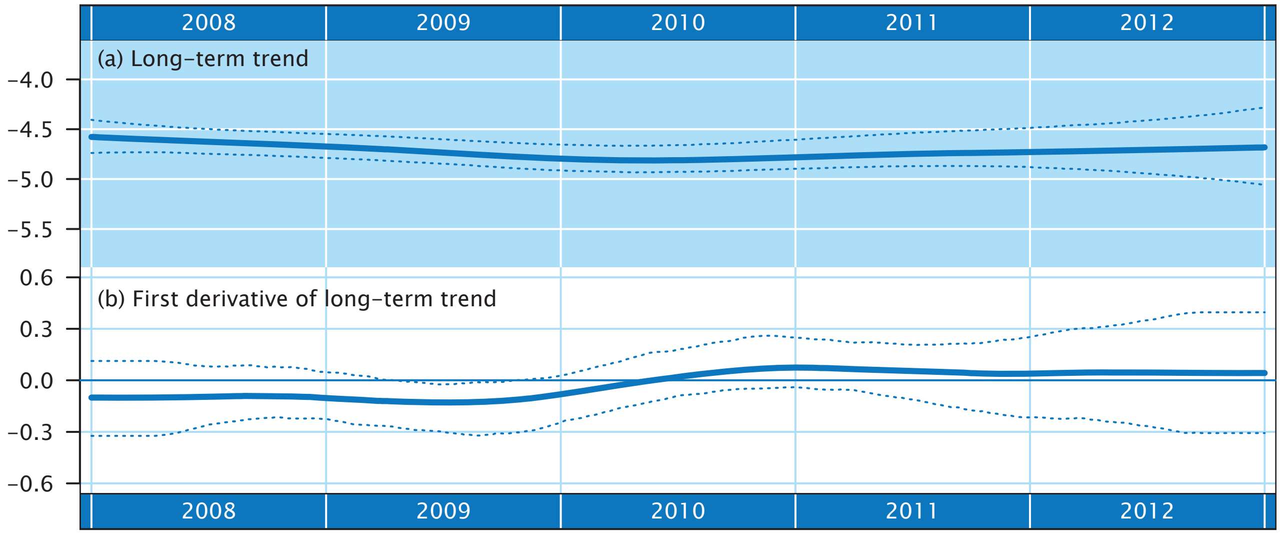
A plot of the estimated annual periodic cycle $f_2(t)$ is shown in Figure 11. The only feature of this cycle that is statistically significant is a slight dip around the month of July; during the rest of the calendar year, the rate of personal crime incidents is not significantly different from the annual average.

Estimated coefficients, standard errors and p-values from this model are shown in Table 3. The coefficient for WLHH is small (−.0014) and not significantly different from zero ($p = .495$)). As we conjectured, the coefficient for Cluster.1 (−.2295) is negative and significant ($p = .000$). Interviewers whose assignments include a higher proportion of housing units in Cluster 1 tend to report fewer incidents of personal crime. The coefficient for After.Realignment is small (−.0527) and insignificant ($p = .614$), so there is no evidence that the field realignment program of 2012 affected the collection of personal crimes.

Figure 10

Long-Term Trend for Incidents of Personal Crime

(a) Estimated long-term trend for rate of personal crime incidents on the log scale (solid line) with pointwise 95% error bounds (dashed lines), and (b) estimated first derivative of the long-term trend (solid line) with pointwise 95% error bounds (dashed lines)



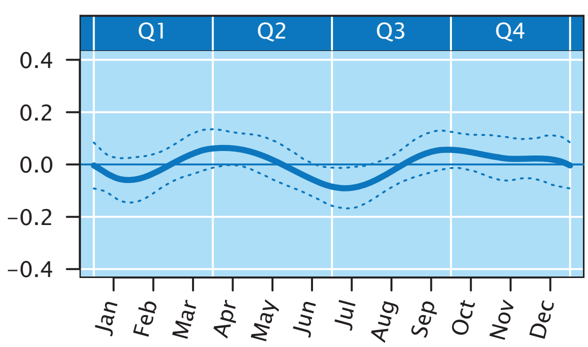
Source: United States Census Bureau, *National Crime Victimization Survey*, 2008–2012

The coefficient for After.Training.2011 is positive (.1459). Although it is not significantly different from zero ($p = .111$), it presents mild evidence that refresher training and performance monitoring increased the rate of personal crime incidents reported in 2011. The coefficient for After.Training.2012, however, is very small (.0138) and insignificant ($p = .897$), so we have essentially no evidence that this program impacted the collection of personal crimes in 2012.

Figure 11

Annual Periodic Cycle for Incidents of Personal Crime

Estimated annual periodic cycle for rate of personal crime incidents on the log scale (solid line) with pointwise 95% error bounds (dashed lines)



Source: United States Census Bureau, *National Crime Victimization Survey*, 2011–2012

In prior analyses for 2011, we concluded that refresher training and performance monitoring did increase the collection of personal crimes that year, but the effect was limited to crimes that were not reported to police [1]. A crime that was reported to police might be more salient in the respondent's memory and more easily discovered by an FR during a screener interview, whereas an unreported crime might be discovered only if the interview is done carefully. One of the major goals of refresher training and performance monitoring was to encourage FRs to conduct the screener

Table 3: Coefficients, standard errors, and Bayesian p-values from model for incidents of personal crime

	Coef	SE	p
WLHH	−.0014	.0021	.495
Cluster.1	−.2295	.0558	.000
After.Training.2011	.1459	.0929	.111
After.Training.2012	.0138	.0980	.897
After.Realignment	−.0527	.1039	.614

Source: United State Census Bureau, *National Crime Victimization Survey*, 2008–2012

Table 4: Coefficients, standard errors and Bayesian p-values from model for incidents of personal crime, classified by by whether the crime was reported to police

	All personal crimes			Reported to police			Not reported to police		
	Coef	SE	p	Coef	SE	p	Coef	SE	p
WLHH	−.0014	.0021	.495	−.0007	.0024	.782	−.0029	.0028	.306
Cluster.1	−.2295	.0558	.000	−.1987	.0663	.003	−.2568	.0757	.001
After.Training.2011	.1459	.0929	.111	.0224	.1178	.832	.2948	.1271	.018
After.Training.2012	.0138	.0980	.897	−.0707	.1201	.581	.0976	.1309	.456
After.Realignment	−.0527	.1309	.614	−.0975	.1242	.423	.0292	.1374	.829

Source: United States Census Bureau, *National Crime Victimization Survey*, 2008–2012

more carefully. Therefore, if this program had an impact on the collection of crimes, it is reasonable to think that the effect would be stronger for crimes that were less salient.

To see if our present analyses support a similar conclusion, we applied the current model (10) just to personal crimes that were reported to police, and again to personal crimes that were not reported to police. Coefficients, standard errors and p-values from these separate models are shown in Table 4. For crimes reported to police, the coefficient for After.Training.2011 is small (.0224) and insignificant ($p = .832$), but for crimes not reported to police, the coefficient is large (.2948) and significant ($p = .018$). Thus, the results for 2011 shown here are highly consistent with what we found in our previous analyses [1].

However, these effects of training and monitoring appear to have been short-lived. The coefficients for After.Training.2012 shown in Table 4 are small and insignificant. The elevation in personal crimes due to training and monitoring seen in late 2011 was apparently not sustained into 2012.

7 RESULTS FOR PROPERTY CRIME

Our final model describes incidents of household property crime. The response variable is

$$Y_{ij} \;=\; \text{household property crimes discovered by interviewer } i \text{ during month } t_j.$$

As in our previous model, we assume

$$Y_{ij} \sim \text{NegBin}(\,\alpha = \kappa^{-1},\, \beta = \kappa^{-1}\mu_{ij}\,),$$

where κ is a dispersion parameter. The offset is

$$\omega_{ij} = \log n_{ij},$$

where n_{ij} is the number of household interviews conducted by interviewer i during month t_j. In all other respects, this model has the same form as the previous one (10). The estimated long-term trend and its first derivative are plotted in Figure 12. A mild but statistically significant decline took place in 2008 and 2009, but from 2010 onward the trend is essentially flat. According to official national estimates published by BJS, however, the rate of property crime victimization rose between 2010 and 2011 and between 2011 and 2012, and both increases were statistically significant [11] [12]. We can think of several possible explanations for this apparent discrepancy. First, the national estimates were weighted to adjust for the sample design, nonresponse and other complications, whereas the model that generated the trends in Figure 12 was applied to raw, unweighted survey responses. Second, the national estimates represent marginal rates, whereas the trends shown in Figure 12 condition upon some time-varying covariates. Third, the estimated curves in Figure 12 contain a fair amount of noise; in particular, the error bounds on the first derivative are wide enough that we cannot rule out the possibility that the rate may have increased from 2010 to 2012. Although the shape of the trends in Figure 12 do not closely mimic the trend in national estimates, the discrepancy is not large enough to be worrisome.

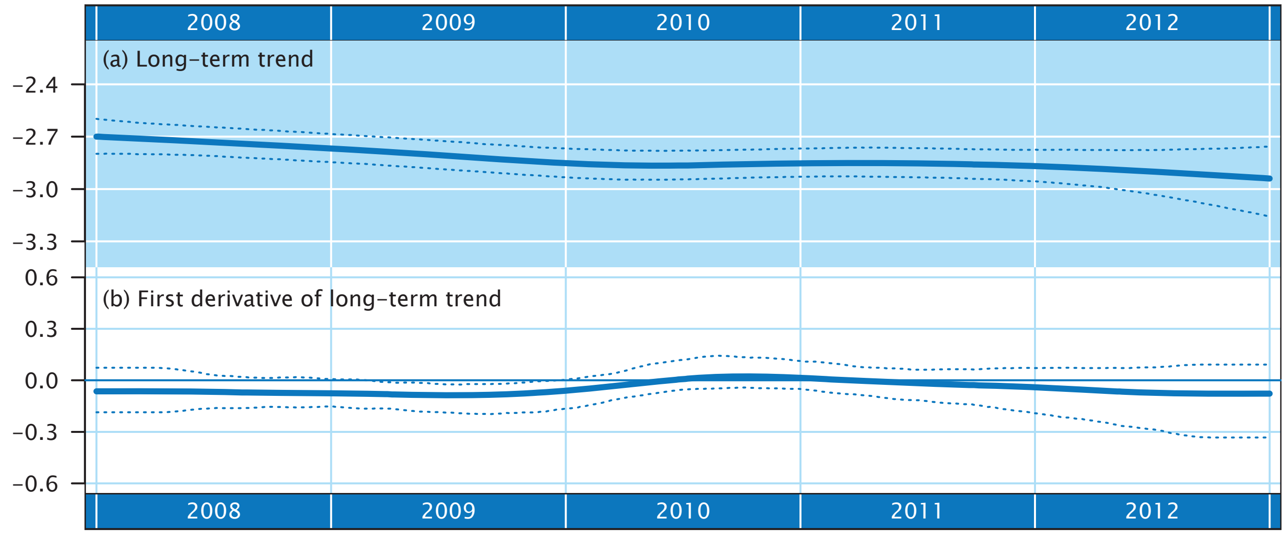
The estimated annual periodic cycle for household property crime is shown in Figure 13. Rates are significantly higher than the annual average during the holiday season of late November through January, and significantly lower than the annual average during May, June and July.

Table 5 shows estimated coefficients, standard errors and p-values from the property-crime model. As in the personal-crime model, the coefficient for WLHH is close to zero (.0022) and insignificant ($p = .086$), whereas the effect of Cluster.1 is negative (−.1907) and highly significant ($p = .000$). As we conjectured, a higher

Figure 12

Long-Term Trend for Incidents of Household Property Crime

(a) Estimated long-term trend for rate of household property crime incidents on the log scale (solid line) with pointwise 95% error bounds (dashed lines), and (b) estimated first derivative of the long-term trend (solid line) with pointwise 95% error bounds (dashed lines)



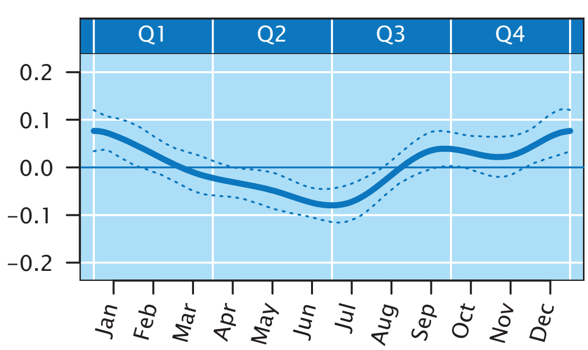
Source: United States Census Bureau, *National Crime Victimization Survey*, 2008–2012

proportion of households in Cluster 1 is associated with lower rates of property crime. The coefficient for After.Realignment is very small (−.0089) and insignificant ($p = .834$), so there is no evidence that the collection of property crimes was impacted by the field realignment. Refresher training and performance monitoring led to a significant increase in the collection of property crimes in 2011 ($p = .012$) but not 2012 ($p = .232$).

Figure 13

Annual Periodic Cycle for Incidents of Property Crime

Estimated annual periodic cycle for rate of household property crime incidents on the log scale (solid line) with pointwise 95% error bounds (dashed lines)



Source: United States Census Bureau, *National Crime Victimization Survey*, 2011–2012

In previous analyses from 2011, we found that the effects of refresher training and performance monitoring on property crimes were moderated by whether the crimes were reported to police [1]. To see if this pattern appears in the current analysis, we re-fit the property-crimes model seperately to crimes reported to police and crimes not reported to police. Results from these models, which are summarized in Table 6, show a similar pattern to what we found with personal crime. In 2011, refresher training and performance monitoring

Table 5: Coefficients, standard errors, and Bayesian p-values from model for incidents of property crime

	Coef	SE	p
WLHH	.0022	.0013	.086
Cluster.1	−.1907	.0332	.000
After.Training.2011	.1196	.0473	.012
After.Training.2012	.0624	.0525	.232
After.Realignment	−.0089	.0574	.834

Source: United State Census Bureau, *National Crime Victimization Survey*, 2008–2012

Table 6: Coefficients, standard errors and Bayesian p-values from model for incidents of property crime, classified by by whether the crime was reported to police

	All property crimes			Reported to police			Not reported to police		
	Coef	SE	*p*	Coef	SE	*p*	Coef	SE	*p*
WLHH	.0022	.0013	.086	.0009	.0017	.614	.0025	.0017	.119
Cluster.1	−.1907	.0332	.000	−.1125	.0447	.008	−.2464	.0404	.000
After.Training.2011	.1196	.0473	.012	.0845	.0708	.231	.1466	.0605	.017
After.Training.2012	.0624	.0525	.232	.0835	.0823	.284	.0474	.0620	.448
After.Realignment	−.0089	.0574	.834	−.0062	.0868	.898	.0060	.0692	.946

Source: United States Census Bureau, *National Crime Victimization Survey*, 2008–2012

increased the collection of unreported property crimes but had no discernible effect on reported crimes. In 2012, the effects of training and monitoring were small and insignificant.

8 IMPLICATIONS

Although our previous analyses showed some effects of refresher training and enhanced monitoring on collection of crimes by interviewers who had experienced that intervention in late 2011 [1], the present analyses show that those effects did not continue into 2012. Using longitudinal models that account for long-term trends, periodic cycles, other covariates, and interviewer variability, we could not detect any significant intervention effects at all on personal or household property crime in in 2012.

The estimated victimization rates from the 2011 survey were not affected by the training and monitoring intervention, because the published estimates were based only on pre-intervention interviews. Because no significant effects on crime have been detected in 2012, there is no evidence to suggest that comparisons between 2012 and previous years whave been impacted by the recent field interventions.

ACKNOWLEDGMENTS

Mike Planty, Jennifer Truman and Lynn Langton at BJS helped to guide this research. Meagan Wilson supplied an overview of refresher training and performance monitoring. Useful suggestions were provided by William Samples, Dave Watt, Tim Kennel, Tammy Adams, Tom Louis, Derek Young, James Livsey, and members of the NCVS Data Review Panel. Special thanks to David Watt, Tammy Adams and Doug Galagate for their generous help in creating and managing the required datasets.

REFERENCES

[1] Schafer, J.L. (2013) Effects of interviewer refresher training and performance monitoring in the 2011 National Crime Victimization Survey. Center for Statistical Research & Methodology Research Report Series (Statistics #2013-07). Washington, DC: U.S. Census Bureau.

[2] 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.

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

[4] White, I.M.S., Cullis, B.R., Gilmour, A.R. and Thompson, R. (1998) Smoothing of biological data with splines. In *Proceedings of the International Biometrics Conference*, 1998.

[5] Welham, S.J. (2008) Smoothing spline models with longitudinal data. In G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs (Eds.), *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*, 253–289. Hoboken, NJ: Chapman and Hall/CRC Press.

[6] Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: SIAM.

[7] Green, P.J. and Silverman, B.W. (1994) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* London: Chapman and Hall.

[8] Eilers, P.H.C. and Marx, B.D. (1996) Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89–121.

[9] Ruppert, D., Wand, M.P. and Carrol, R.J. (2003) *Semiparametric Regression.* New York: Cambridge University Press.

[10] Zhang, D., Lin, X. and Sowers, M.F. (2000) Semiparametric regression for periodic longitudinal hormone data from multiple menstrual cycles. *Biometrics*, 56, 31–39.

[11] Truman, J., Langton, L. and Planty, M. (2013) *Criminal Victimization, 2012.* U.S. Department of Justice, Bureau of Justice Statistics, NCJ 243389, October 2013.

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