FINAL RESULTS OF SAMPLING PLAN FOR THE PHILADELPHIA POLICE DEPARTMENT QUALITY ASSURANCE BOARD (QAB)[1]

Submitted by

Susan Kinnevy, Ph.D.
Neil A. Weiner, Ph.D.
Vera Huang, MS
School of Social Work
Center for Research on Youth and Social Policy (CRYSP)

and

Elaine Zanutto, Ph.D.
Department of Statistics
University of Pennsylvania

Submitted to

The Philadelphia Police Department

October 2004

[1] Funding provided by the Philadelphia Police Department.

Table of Contents

Project Overview

The major task for the project was the development of a computerized sampling plan and a procedure for implementing the protocol in the field to enable QAB to use the field-tested sampling techniques in their review of Philadelphia Police Department incident and summary statistical reports. Identifying error rates through rigorous statistical techniques is important in several ways:

Deployment Decisions

Demands for deployment of personnel are partially a result of the numbers and types of crimes in different districts of the city. Unless the data on which these decisions are made are accurate and relatively free of error, the decisions will not optimally deploy personnel in a strategic manner. The use of statistically rigorous sampling techniques, and an analysis and understanding of the error rates generated by the implementation of these techniques, can inform equally rigorous deployment decisions.

Civil Liability

Although there are crime classification errors of various types, certain types are more critical than others. For instance, the classification of either Part I or Part II crimes as Other, i.e. non-crimes, not only does a disservice to the victims involved, but may also place the Department at risk of civil liability action.

Patterns and Trends from a National Perspective

Unless crime report data are reliable and valid, the patterns and trends in Philadelphia are more difficult to detect. Not only does this create problems for understanding municipal department performance, it also affects the way in which Philadelphia is ranked nationally according to the Uniform Crime Report (UCR)

Program. In addition to its importance as a crime measure, the UCR is often used as an indicator of quality of life, along with other social and economic data.

In view of the importance of accurate crime data, the current project is critical because it moves the Department in a direction designed to promote and ensure that accuracy. In the following report, we (1) summarize project activities along a timeline; (2) introduce terminology used in the report, and (3) present the results of the implemented sampling plan to detect errors in coding in incident reports.

Project Timeline

11/02- 12/02:	CRYSP met with QAB staff regarding their current sampling plan and their needs in the future.
1/03:	CRYSP worked with a sampling expert from Penn's Wharton School to determine the best sampling procedure and to develop the equation (error rate calculation Excel table) to apply to the sample.
2/03:	CRYSP delivered sampling plan to QAB staff; staff agreed to the plan; CRYSP developed methodology for sample selection.
3/03 -6/03:	CRYSP worked with QAB to implement the sampling procedure in the field for reports from January through April 2003 (two-month time lag between reports and sampling).
7/03:	QAB supplied CRYSP with data on all reports from January through April 2003 to allow for data analysis and error rate calculations.
8/03- 11/03:	CRYSP calculated error rates based on a 4-month sample and analyzed results.
12/03:	CRYSP prepared an Interim Report based on the 4-month sample, but felt that the error rates would be more stable if calculated using data for 12 months. CRYSP requested this data from QAB.
7/04:	QAB supplied CRYSP will data on all reports from January through December 2003.

9/04:	CRYSP Final Report, including a CD with the error rate calculation table that contains the equation to be applied to monthly samples.

Terminology in Report

Following is a set of definitions for the terminology used in the data analysis. These definitions should be helpful in the interpretation of results.

Random Sampling: A procedure for selecting a subset (sample) of incident reports from a larger set (population) in a way to ensure that the sample reflects the composition of the population in areas of interest (e.g., with respect to crime coding errors). In this report we used a special kind of random sampling method in which each bundle of incident reports had an equal chance of being selected.

Recoding: Changing an incident report's crime code designation to another crime code designation.

Reclassification: Changing an incident report's crime-code category to another category (e.g., Part I to Part II, or Part II to Other).

Error Rate: The percentage of miscoded or misclassified incident reports, whether by district, crime type, or citywide.

Standard Error (SE): A measure of how accurately the sample-based error rate reflects the error rate in the population. The smaller the standard error, the better the error rate statistic reflects the error rate in the population.

Confidence Interval, Lower Limit, and Upper Limit: The confidence interval represents the range of error rates within which the actual population error is likely to fall. The lowest of these error rates is called the "lower limit," and the greatest of these

error rates is called the "upper limit." The sample-based error rate typically falls in the middle of the interval.

Statistically Significant Difference: Statistical significance refers to the probability that one error rate in the sample (e.g., for robbery) is significantly different from another error rate in the sample (e.g., for aggravated assault) or from some specific value of interest, such as zero.

The terms defined above are explained further in the Expanded Glossary (Appendix E) which provides additional technical details, as well as in the following analysis and interpretation of the results of the sampling implementation.

Results: Implemented Sampling Plan to Detect Coding Errors in Incident Reports

The following tables and charts represent findings from our analysis of review data for the sampling period January through December 2003. The sampling plan was based on the random sampling of two and three-day bundles of reports, by month and by district. There were 1,838,254 incident reports in 2003. The sampled bundles yielded 113,742 incident reports that were reviewed (6.19%) from among the total 1,838,254 reports, of which 1,677 (1.47%) were found to be in error and either recoded or reclassified. Although this percentage of error is not very large, it is statistically different from zero. Some districts and crimes types contribute coding errors systematically and disproportionately to this percentage.

Of those reports that were found to be in error, 64.7% were recoded or reclassified as more serious crimes. As there was no record kept of which reports were reviewed and not reclassified, CRYSP was unable to compare the reports that were in error with the reports that were not in error.

Table 1 shows the distribution of crime code errors and how they were reclassified. Of the 1,677 reports found to be in error, 283 were from Part I reports, 631 were from Part II reports, and 763 were from Other reports. Of the 1,667 errors, 1,078 (64.7%) were reclassified as more serious crimes (Part II to Part I, or Other to either Part II or Part I, *marked in red in the table*). Within Part I report errors, 229 were recoded as different Part I crimes, 35 were reclassified as Part II crimes, and 19 were reclassified as Other crimes. Within Part II report errors, 325 were reclassified as Part I crimes, 261 were recoded as different Part II crimes, and 45 were reclassified as Other crimes. Within the Other report errors, 192 were reclassified as Part I crimes, 561 were reclassified as Part II crimes, and 10 were recoded as different Other crimes. When tested for statistical significance, this pattern in recoded and reclassified errors was found to be statistically significant, which means that the recoding and reclassification did *not* form a random pattern. In other words, the probability that an incident report would be recoded or reclassified in a particular manner depended upon the crime-code class in which it originally fell. Basically, Table 1 demonstrates that if an incident report was in a *less serious* crime-code class, it was more likely to be reclassified to a *more serious* crime class. Put differently, an incident's seriousness was more likely to be *under-recorded* than over-recorded.

Table 2 shows the citywide error rates for Part I crimes individually, and for Part II and Other crimes in the aggregate. The error rates are rank-ordered from high to low, with larceny yielding the highest (3.0%) compared with other crime types and with the citywide total (1.5%).

Table 3 shows the overall error rate by district in decreasing rank order. Although District 77 had the highest error rate (3.89%) compared with other districts and with the citywide error rate, the district sample also contained the largest standard error and confidence interval, making it more difficult to interpret the degree to which the error rate in the sample represents the error rate in the population of District 77 reports. There are 8 districts where the error rates are significantly different than the citywide error rate. This means that their confidence interval's upper and lower limits do not overlap with citywide upper and lower limits. This approach to inferring "statistical significance" will be illustrated graphically later in this report. Of the 8 districts, four are significantly higher than the citywide error rate (Districts 77, 9, 15 and 22) and four are significantly lower (Districts 17, 39, 19 and 92).

Table 4 shows the distribution of the 631 Part II incident reports found in error and reclassified. Over half of these reports (325) were reclassified as Part I crimes, which could be a cause for concern because of their migration upward to more serious crime codes.

As **Figure 1** illustrates, the largest percentage (61%) of the 325 Part II crimes reclassified as Part I crimes were reclassified as Aggravated Assaults. This skewed finding may indicate that these reclassifications need to be analyzed further to determine if there is a systematic source of this error. **Table 5** shows the distribution of the 763 Other reports found in error and reclassified. In this case, almost 99% of these reports were recoded as more serious crimes (i.e., Part I or Part II).

As **Figure 2** illustrates, the largest percentage of the 192 Other errors reclassified as Part I crimes were recoded as Larceny (54.69%). Insofar as these reports were markedly upgraded in their seriousness, they may merit intensive scrutiny.

Figure 3 illustrates that the largest percentage of the 561 Other errors (73.3%) reclassified as Part II crimes were recoded as Other Assault. This finding may indicate that the distinction between certain subtypes of Other and Other Assault is not clear.

Table 6 shows the distribution of the 210 larceny error reclassifications. Of those reports, 84% were recoded as other Part I crimes, which involves no migration across classification categories**. Table 7** focuses on larceny recoding by crime type. As the table shows, the largest percentage of larceny errors were recoded from one larceny crime to another larceny crime (59.05%), indicating that the relatively large error rate may not be a cause for concern because the shift is not to a more serious crime type.

Looking at the larceny reclassifications in a different way, **Figure 4** illustrates the reclassification of larceny errors by district, specifying the top 13 districts and aggregating the remaining districts. It may be useful in the future to look into the reasons for the concentration of more than 50% of the larceny errors in the top five districts. In order to illustrate how errors are distributed within a district, **Table 8** shows the pattern in the reclassification of the 98 errors in District 22. Well over half the reclassifications (61.23%) were from less serious to more serious crimes, with the largest percentage (23.47%) being a reclassification from Part II to Part I crimes.

Figure 5 illustrates the recoding pattern in District 22, with over 40% of the district's errors recoded into Other Assault or Aggravated Assault. Slightly more than one-fifth (21.4%) of the errors were recoded as Larceny.

To understand how important and how much attention should be paid to an error rate, it is necessary to consider things in addition to the size of the error rate itself. It is also necessary to assess: (1) the degree of confidence we have in the accuracy of the error rate and (2) the error rate's statistical significance. **Figure 6** illustrates the importance of the confidence interval width when examining districts with similar error rates. Districts 3, 8 and 25 have error rates that differ by only a few decimal points. The confidence interval widths (i.e., *the blue bar to the right of each district's lower and upper limits*) are the same in Districts 8 and 25 (.01) indicating that the error rate similarity can be trusted as accurate. The error rate in District 3 is not much different than those in the other two Districts. However, the confidence interval width (.02) is one percentage point larger, indicating that our confidence in the accuracy of this error rate should be less.

With respect to the significance of differences in error rates, it is important to assess not only differences in confidence interval widths, but also where the upper and lower limits of those interval widths fall.

Summary and Conclusions

Summary

- *Bundles Sampled:* Of the approximately 4,500 two- and three-day incident report bundles generated by the 25 districts in 2003, 826 were randomly sampled (18.0%).
- *Incident Reports Yielded for Review:* The 826 sampled bundles yielded 113,742 that were reviewed for coding and reclassification errors, which was 6.19% of the 1,838,254 incident reports citywide for 2003.

- *Citywide Errors:* Overall, 1,677 of the 113,742 reviewed incident reports were found to be in error—for an estimated citywide error rate of 1.5%— and were recoded or reclassified.
- *Direction of Coding Errors:* Of the 1,677 reports that were found in error, 1,078 (64.7%) were recoded or reclassified as more serious crimes (Part II to Part I, or Other to either Part II or Part I).
- *Citywide Error Rates by Crime Types and Crime-Type Groups:* There was noteworthy variation in error rates by crime type. Among the eight Part I crimes, larceny and rape had the highest error rates citywide (3.0% and 1.9%, respectively), and arson and homicide the lowest (.8% and .4%, respectively). Part II crimes had the highest error rate (3.2%), followed by Part I (1.9%) and Other (1.0%).
- *District Error Rates:* There was also noteworthy variation in error rates by district. Among the 25 districts, district 77 had the highest overall error rate (3.89%) and district 92 had the lowest (.29%). The overall error rates in Districts 77, 9, 15, and 26 were significantly above the citywide total error rate, and the overall error rates in districts 17, 39, 19, and 92 were significantly below the citywide total error rate.
- *Reclassification of Part II to Part I:* Among the Part II crimes reclassified as Part I crimes, three-in-five were reclassified as Aggravated Assaults.

- *Reclassification of Other to Part I:* Among the Other crimes reclassified as Part I crimes, one-in-two were reclassified as larcenies.
- *Reclassification of Other to Part II:* Among the Other crimes reclassified as Part II crimes, three-in-four were reclassified as Other Assaults.

Conclusions

- *Magnitude of the Error Rate:* Regardless of the district and the crime type, between one and two (1.5%) incident reports among every 100 written were in error. The error rate may seem low, but its substantive importance depends upon how it is used or its impact. Personnel deployment decisions, civil liability burdens, and comparative national benchmarking would invite quite different interpretations of whether an error rate of 1.5% is low or high.
- *Variability in the Error Rate:* The error rate varies by crime type and by district. Remedial activities, for example, whether training or technical assistance, can be geared to those crime types and districts that are most error prone in order to obtain the most potential positive change.
- *Criteria for Determining Corrective Action:* Statistical significance is a possible tool for making decisions about which crime types and districts most merit corrective action. Other criteria might include public perceptions, crime severity, and issues of civil liability and personnel deployment.

- *Reasons for Recoding and Reclassification:* More work needs to be done to focus on the reasons incident reports were recoded. In particular, a record needs to be kept of incident reports that were reviewed but not reclassified in order to understand reasons for recoding.

Recommendations

Continuous Quality Control

The Department should consider how often to review the error rates generated through the sampling design and the levels of error that are acceptable. Based on these assessments, the Department should further consider what remedial corrective actions are required.

Shift in Sampling Design

The ideal revision in the sampling would be to shift from bundle sampling to simple random sampling. This shift will reduce substantially the number of incident reports that need to be reviewed in order to achieve the appropriately small confidence intervals (see Appendix D). As we understand it, Part I crime reports are currently available online for review because they are scanned in as pdf files. If the Department agrees to expand the scanning to include Part II and Other crime reports, the shift will be easy to accommodate. If this shift proves too cumbersome, it may be possible to at least scan in Part II and Other crimes reports on an ad hoc basis for special QAB projects.

Implement External Audit

Regardless of which sampling design is implemented, an independent audit of the reports reviewed during the first quarter of the year (January – March, 2004) should be

conducted to compare outcomes with the QAB review for that same time period. Analysis of outcomes from the QAB review and the independent audit may yield a set of suggested revisions/adaptations to whichever sampling protocol has been implemented. At that point, a schedule for independent audits for the remainder of the year should be determined. The independent audit is important, then, for two reasons:

1. To inform the direction that the sampling design takes in the future.
2. To determine the manner and number of audited reports that are required to remain confident about the quality of the error reporting.