An Experiment with Process Control: Implementation and Results

Claudia S. Glover, Daniel L. Whitehouse, Charles H. Alexander

I. INTRODUCTION

For its survey and census operations, the Census Bureau routinely uses various acceptance sampling procedures to control quality of the product of each operation. The quality control procedures for the current procedures for the Current Population Survey are summarized in Brooks and Bailar (1978); the CPS procedures are generally representative of the Bureau's quality control methods.

In the past few years, the Census Bureau has been seeking ways to apply new methods of "process control" to improve its operations, including the approaches to quality and productivity described in Deming (1982). A software system for providing process control information, based on some of these approaches has been developed at the Census Bureau. This system, known as the Data Quality Assurance Software System (DQASS), is described in Diskin (1988). This paper describes the results of an experiment to test the effect of implementing a process control approach using this software for one of our clerical operations.

In the experiment, a coding and transcription task from the National Longitudinal Mortality Study (NLMS) was divided between two independent but similar clerical staffs. One group used the Census Bureau's usual quality control procedure and the other used process control techniques modelled after ideas in Deming (1982). Error rates and production rates were measured for both groups of clerks. An independent reverification to measure outgoing error rates for a sample of the work was done by one of the authors. These numerical results are presented below, along with some qualitative observations by Census Bureau staff who observed the experiment. Further details and some recommendations for future Census Bureau operations are included in Glover, Whitehouse, and Alexander (1988).

For the NLMS, data are coded and transcribed from death certificates ordered from each state. The data include death certificate numbers, name, sex, ethnicity, race, date of birth, place of death, time of death, cause of death, and whether or not the death occurred in a hospital.

In all, 22 items were coded or transcribed from each death certificate. Thirteen items required coding, 6 were straight transcription, and the other 3 could require either coding or transcription. The operation is complicated by the fact that each state has its own deatl certificate form, with some differences in the formats and the available information.

II. IMPLEMENTATION OF THE EXPERIMENT

A total of 19,958 death certificates were included in the study. The death certificates were processed one state at a time. They were divided into "work units" consisting of about 50 certificates.

Death certificates were randomly assigned to either the control or experimental group. The certificates were stratified by state and within state by year of death. The control and experimental groups used acceptance sampling and process control methods, respectively. Each group had a supervisor, lead clerk and 7 coders. The coders in the two groups were requested to have comparable work experience, be similarly recruited and have the same Civil Service grade. However, the coders were not randomly assigned to the two groups. The experiment was designed such that the only deliberate differences between the two groups were the proportion of death certificates which they verified, the training of the supervisor, and the type of feedback they were given. The control group dependently verified 100% of the death certificates which they coded and transcribed and (after the first week) the experimental group dependently verified 25% of the death certificates which they coded and transcribed. Both groups received feedback of their performance. The supervisor of the control group used only the individual verification records to give feedback to her group and to monitor their performance. The supervisor of the experimental group was provided computer-generated data from the DQASS system. This supervisor was given special training in using this information for process control.

The main uutput from the DQASS is as follows:

1. Control Chart. The control chart is a graph of error rates (or other measure of performance) over time for each coder. The error rate for each work unit was graphed. The control chart consisted of a process average along with upper and lower control limits which, for this experiment, were ±2 standard deviations from the process average.

2. A list of data consisting of: Work units out of control; coders with runs and trends; and coders whom the DQASS identified as needing feedback.

3. Error Matrices. Two types of error matrices were provided:
   a. Distribution of errors by coder and type:
      1. Incorrect transcription
      2. Failure to transcribe
      3. Incorrect code
      4. Failure to code
      5. Illegible (no charge error)

b. Distribution of errors by item and coder

III. LIMITATIONS ON INFERENCES FROM THE EXPERIMENT

Because the experiment had to be fit into a production operation, there were some departures from strict principles of experimental design. These departures, along with some other potential confounding effects, are discussed in this section.

A. Confounding of Supervisor Effect with Experimental Treatment.

The control group and experimental group had different supervisors and lead clerks who served as assistant supervisors. Thus, it is difficult to distinguish between differences due to the method of quality control and inherent differences between the supervisors. The experiment measures the differences between the two methods as used by these particular supervisors. No statistical inferences can be made to a larger population of supervisors. For this reason, it was important to observe as much as possible about how the process control information was used. Observations of this kind are found in a subsequent section.

B. Lack of Random Assignment of Coders.

The control group was selected to be from the usual branch which did this coding/transcription operation. It would have been administratively awkward to randomly assign clerks to this branch specifically for the experiment. Instead, an experimental group was formed by judgmental selection from the same pool of potential applicants as was used for the control group. The coders in the two groups were judged to be generally similar in experience and background. However, no specific data on coder experience or background were collected.

C. Possible Differences in Training of Coders.

The normal procedure is for the supervisors to train the coders. This introduces an additional possible difference between the two groups. Standardized training by someone other than the supervisor was considered, but there was not time prior to the experiment to develop a training procedure which we were sure was as effective as the usual training. Since different supervisors were unavoidable during the operation, the decision was to let each supervisor do as well as possible using her usual training techniques.

The groups were trained separately. Each branch was instructed to use its usual training pro-cedures. Some differences in the training were observed. The supervisor of the control group gave each coder a Coding/Transcription Operation procedure, had them read and study it, then reviewed and discussed it with the group. The supervisor of the experimental group gave each coder a Coding/ Transcription Operation Proce-dure, read aloud the procedure, discussed it with the group, coded and transcribed a sample death certificate with the group, then gave each of the coders 5 sample death certificates to code and transcribe. When everyone finished, she reviewed and discussed the answers with the group.

D. Possibility of Interaction of the Groups.

Both groups of coders knew that an experiment was going on, but an attempt was made to downplay the importance. The coders were instructed not to discuss the operations with members of the other group. The control group was told only that the other group would be receiving special feedback. No penalties or awards were given based on the performance of either group.

Several considerations suggest that communication between the group probably had little influence on the results:

1. The two groups were located in separate buildings and had no contact during working hours;

2. Details of clerical procedures are not thought to be a common after-hours topic of conversation among the clerical staff.

3. Observers noted no signs of interaction.

4. Even if the coders in the control group knew what kind of feedback the experimental group was receiving, they did not actually receive such feedback themselves.

E. Unintended Difference in Coders

It was intended that both groups of coders be recruited from the same Civil Service register, the Competitive Register. During the experiment, we discovered that one of the control group coders had been hired from a different register, the Handicapped Register. To try to limit the possible effect of this difference, starting on Monday, June 29th, we decided not to have this coder verify other coder's work. Some differences were observed for this coder at the start of the experiment.

At the end of the first week of the experiment, her production rate was 1/hour while the group's production rate was 10/hour. Also, her error rate was 4.42% while the group's error rate was 1.46%. However, by the end of the experiment the coder had moved toward the middle of the group. The coder's overall average error rate and production rates were .67% and 5/hour, respectively. Her error rate was the fourth lowest in the group.

F. Difference in Continuity of Supervision

As mentioned previously, each group had a supervisor and lead clerk who also had some responsibilities for other projects. Most of the time at least one was available to answer questions and give feedback. There were no specific protocols concerning how much time each one was to spend in front of the group. This led to some differences between the two groups.

In general, the experimental group tended to use the lead clerk to concentrate on other projects while the supervisor concentrated on the NLMS operation and experiment. For the control group, there was comparatively more equal division of time. It is not clear which, if either, strategy is preferable. Greater concentration by the supervisor may facilitate more intense feedback, but it may create problems when the supervisor has a day off. The experimental group's supervisor had two days off during the experiment.

G. Possible Effect of Observers

The previous NDI coding/transcription operation was observed for two days by the project manager, who served as a "subject matter" expert, to be sure that the coding instructions were followed correctly. In this experiment, the project manager also served as observer of the feedback in both groups, and was present during most of the experiment.

The presence of the project manager in observing and monitoring the experiment and her presence as the subject matter expert for both groups may have had a conscious or unconscious effect on both groups' perception of the operation, thereby affecting their quality and production. However, the project manager did not interact directly with the coders.

H. Implementation Difficulties

The DQASS software had been used for previous projects, and it needed only relatively minor modifications to produce the output needed for this project. Unfortunately these modifications were not completed sufficiently far in advance of the experiment; indeed some modification and testing was still going on as the experiment began. Besides producing a sense of panic, this could well have reduced the effectiveness of the feedback.

One obvious problem was that there was little time to work with the experimental group supervisor on how best to use the DQASS information to give feedback. Fortunately the supervisor was able to develop an effective style of doing this. Also, the supervisor had to make some modifications to the DQASS make it more convenient for her use. This is described in the qualitative observations section below.

IV. QUALITY ASSURANCE TECHNIQUES USED

The notion of a limited experimental application of Deming's quality improvement principles is something of a contradiction in terms, since Deming's philosophy requires a complete commitment to quality improvement. Only a few of the changes recommended in Deming (1982) were incorporated into this experiment.

The control group used 100% verification, as is standard for such DPD operations. The coders verified each others' work. If errors were detected, they were corrected. The supervisor reviewed individual verification records to see what errors were made and discussed these errors with the coders.

The experimental group used 100% verification for the first week and 25% verification thereafter, after the control limits were determined. As with the control group, the coders verified each others' work. If errors were detected during the verification, they were corrected. In addition to the individual verification records, the supervisor had at her disposal error matrices showing the type of error by item in error for each coder and for the entire group, and control charts showing error rates as a function of time for each coder and for the entire group. The supervisor was trained and encouraged to use this information to detect problems in the system and to make improvements.

The supervisors of both groups were given considerable discretion in how to use the available information. This was primarily because we were not sure what approaches would be most effective, and we believed we could learn from observing what uses the supervisors chose to make of the different information.

The type of error and item in error information in the error matrices and the individual verification records proved to be extremely useful for spotting problems in the NDI coding process. Coding for the NDI is relatively complex. Failure to understand how the codes are to be assigned in special circumstances was a frequent cause of errors, especially when a particular state's death certificates had a confusing format. Specific errors made by one coder frequently alerted supervisors to a potential major misunder standing which needed to be explained to the entire group.

An error matrix showing type of error crossed with item in error was considered, but not used. For the NDI there are only two common types of errors and for most items one type of error is likely. Therefore, such a matrix would not add much information. For other studies, this kind of matrix might be much more useful.

The control charts proved to be of less value than expected. This was primarily because the groups were so small (and the error rates so low) that the supervisors tended to review almost all errors, even in the experimental group. However, special attention was given in cases when a coder whose work was previously in control exceeded the upper control limit. The control charts might have been much more useful for a project with a larger clerk-to-supervisor ratio, where the supervisor would have to be more selective about which clerks would get the feedback each day.

A second problem with the control charts was observed for this particular study. For coders with low error rates, the normality assumptions implicit in the usual control chart analysis were sometimes violated. This problem mainly occurred for the sample verification, which reduced the number of death certificates associated with each point on the control chart. For some individual coders, zero or one error put the error rate within the limits, while two errors was sufficient to place the coder out of control. Use of 3 standard errors instead of 2 standard errors might have reduced this problem.

Some aspects of the Census Bureau's traditional acceptance approach were incorporated in the protocols for both groups. Specifically:

1. During sample verification, if the experimental group had work units with error rates equal to or above 3%, the work units were to be rejected and dependently 100% verified. The control group was not subject to this rule because their work was being 100% verified.
2. If a coder from the control group or experimental group had three consecutive work units out of 10 work units with error rates equal to or above 3%, the coder was to be retrained.

Since none of these conditions ever was encountered by the experimental group, these conditions probably had a negligible effect on that group.

V. QUANTITATIVE RESULTS

A. Error Rates From the Verification Process

The average error rate for the entire operation was 0.61% for the control group as measured by the verifica tion process. The corresponding rate for the experimental group was 0.31%, about half the rate of the control group.

It is questionable whether this difference can be ascribed to the DQASS feedback, however. A difference between the two groups was already observable after the first week: 1.12% for the control group vs 0.58% for the experimental group. The feedback process based on the automated data was not well underway during much of that week.

The percentage of errors which were transcription rather than coding was about the same for both groups: 27.2% for control vs 26.5% for experimental. The control group did have a higher percentage of items left blank ("failure to code or transcribe") than the experimental group: 4.7% vs 2.1%. We do not know the reason for this difference.

Some differences were observed concerning which items were most likely to be in error. The top two items from the control group, the relationship of the informant to the decedent and the interval between the onset of the illness and the immediate cause of death, were substantially more common than the next most common error. The top items were less salient for the experimental group, i.e., the experimental group seemed to have a more even distribution of items in error.

The most common items in error are all classified as coding items, suggesting that the items were difficult to code, or that there may have been a problem with the coding procedure or its interpretation.

B. Rates of Production

The two groups completed their work on exactly the same day. The control group coded somewhat faster: 13.9 certificates per hour compared to 11.8 for the experimental group. As expected, the experimental group spent fewer hours verifying, 248 compared to 434.

The experimental group verified at a much lower rate than the control group. Some differences might be expected because of additional "overhead" in conducting the sample verification. However, the large difference may suggest that the experimental group was more conscious of the experiment and spent an inordinate amount of time carefully coding and verifying the work, or that the control group coded and verified at an unusually fast rate.

The duration of the project was about half as many weeks as was expected based on a previous NDI coding and transcription operation. In the previous operation, the production rate was about 7 deaths certificates per hour.

C. Analysis of Individual Coders

Error rates and production rates for individual coders were analyzed. There was a high rank correlation between the first week and the final error rates, but a relatively low correlation between production rates and error rates.

D. Results of Independent Reverification

The verification used during the operation was dependent, i.e., the verifier looked at the original transcription sheet while doing the verification. Also, the verification was done by other members of the group which had done the original coding. In general, dependent verification is expected to miss more errors than a completely independent verification.

An independent reverification was conducted for a small sample of the death certificates, about 200 from each group, to measure the actual outgoing error rate. The reverification was performed by a subject matter expert (the project manager).

The results of the reverification showed an outgoing error rate of 0.82% for the control group and 0.75% for the experimental group. These rates were not significantly different.

If the original verification had been independent, one would have expected essentially no errors in the control group. For the experimental group, in which 89% of the records were verified at a 25% rate, with the verification giving an error rate of 0.31%, one would have expected a 0.31 x .75 x .89 = 0.21% outgoing error rate. Thus, the results of the reverification indicated that a substantial number of errors escaped the initial verification in both groups.

Based on the above calculations, the estimate would be that more errors were missed in the control group (0.82%) than in the experimental group (0.75% - 0.21% = 0.54%). However, because of the relatively small number of cases in the reverification, and the variance of all the estimates which go into these calculations, this difference is not statistically significant.

About 40% of the outgoing errors from both groups results from the incorrect coding of one particular item, patient status of decedent at time of death (i.e., Dead on Arrival at hospital, tient in hospital, outpatient/ emergency, etc.) This was not one of the most common items in error as measured in the original verification. Therefore, it appears that the coding of this item was not well understood by either the coders or verifiers in either group.

VII. Qualitative Observations

Observers were present to observe the two groups during much of the coding operation. The observers felt that the supervisors and clerks in both groups performed very well. Both groups received feedback of their performance; however, the experimental group received "special feedback."

The feedback procedures actually used in the control group were observed to be the standard ones used for such operations. The supervisor of the control group made frequent use of the verification records to give feedback to her group about specific errors which were noted. This feedback seemed to be presented effectively. The supervisor reviewed the records for accuracy before showing specific errors to the coders.

The supervisor of the experimental group was observed to make active use of the data from the DQASS. She had to be careful not to divulge data about specific coders to the group. Accordingly, she cut out the control charts for each coder and stapled them in each coder's folder. Since the error matrices had names on them, she reviewed them herself and then shared the necessary information with the affected coder. A typical feedback session occurred almost daily and lasted about 5 minutes. The feedback sessions seemed to be welcomed by the coders. The supervisor discussed with each coder his/her control chart, error rates for most recently completed work units, errors made, and average error and production rates.

There seems to have been some degree of "over-control". The supervisor let the coders know how well they were doing in comparison with the group and informed each coder of the group's average error and production rates. She encouraged the coders to try to decrease their average error rate, even if their performance was better than average. Also, before sending the verification records to SMQCB to be keyed into the computer, she reviewed them for accuracy and sometime showed the coders the specific errors they had made.

Although the error matrices could not be shown to the coders, the supervisor judged them to be a good summary of the types of errors which were made and of the type of items which gave the coders problems.

One of the aims of the "process control" approach is to encourage improvements to be made during the course of the operation. The supervisor of the experimental group initiated and instituted the following improvements during the experiment:

1. Some errors were caused by difficulty in reading the coding instructions. During the operation, the supervisor gave the coders enlarged coding cards. The cards were flat, stiff and contained all of the item descriptions and codes. They were enlarged because some of the coders complained about the small print contained in the procedure. Also since all of the item descriptions and codes were on one card, the coders did not have to turn pages to find information. The Coding cards were used as a supplement to the written procedures.

2. During the operation, the supervisor also provided the coders with a magnifying glass which helped the coders to read or decipher some of the letters and numbers written on the death certificate which were otherwise illegible.

3. One of the coders was discovered to have high error rates in the morning and low error rates in the afternoon; apparently this coder was not a "morning person". The supervisor alerted the coder to this tendency and gave the coder special encouragement with respect to the problem. However, it was not an option to have the coder work only in the afternoon.

4. In numerous instances, the experimental group supervisor determined during the feedback that additional explanations of the coding requirements

were necessary. This occurred especially when a new state presented problems because of peculiarities in the format of the death certificate. In these instances, the supervisor would immediately explain the issue to the entire group. This also occurred in the control group. However, the observers' impression was that such explanations were more common in the experimental group. No qualitative data were collected on the frequency of this feedback.

It is possible that the supervisor would have made some or all of these improvements without the DQASS feedback. Before the operation began, she gave the coders "stand-up binders" to store their Coding/Transcription Operation Procedures. The binders served several purposes: there was a place to store the procedures and they made it more convenient to look up and read information in the procedure.

VII. DISCUSSION OF THE RESULTS

Although both groups had very low initial error rates, the experimental group had lower initial error rates than the control group. This was statistically significant as regards random error due to the assignment of death certificates to the groups. The results of the independent reverification seemed to be somewhat favorable to the experimental treatment, but the differences were not statistically significant.

As described in the previous section, observation of the experiment noted some things which might have led to improved error rates in the experimental group. It was observed that the supervisor of the group did in fact make active use of the DQASS output. The supervisor was unusually oriented towards making improvements (e.g. enlarging copies of the instructions) during the operation, and this may have been partially influenced by the feedback information. The feedback sessions, although brief, led to continual free discussions of the coding instructions between the supervisor and the group. Finally, the supervisor expressed positive reactions to the feedback information and stated that the information had been useful to her.

The comparisons of quality were generally favorable to process control. However, it is not clear how far these results can be generalized. It must be emphasized that this experiment cannot prove that the success of the operation is due to the "process control" method of quality assurance. The experiment was not designed so that this was the only difference between the two "treatments". The groups had different supervisors, and may have had differences in the skill or experience levels of the coders. In addition, both groups were aware that an experiment was going on; this could have subtle unconscious effects on either or both groups.

In general, the experimental procedure does not seem to have resulted in major time gains or losses. The experimental group and control group completed the work in the same number of days. The experimental group took longer on the coding/transcription, but made up the difference by spending less time on verification. The actual experimental feedback process took very little time; this was roughly estimated to have been about 5 minutes per clerk per day. The difference in the coding/transcription times may have been due to spending more time discussing how to do the coding outside of the "formal" feedback sessions. The application of process control proved to be workable in the context of our clerical activities. Observation of the process confirmed that the automated process control information was used effectively during the operation. The experimental results, while not conclusive, were generally favorable to the new procedure. There was no indication of fundamental problems which would rule out future applications.

The potential advantages of the automated output might be greater for projects where the clerk-to-supervisor ratio is higher, since the output could identify which clerks were most in need of feedback or further instruction. However, in this case it is important to train the supervisors to be more selective in their corrective actions.

References

Brooks, C.A. and Bailar, W.A. (1978). An Error Profile: Employment as Measured by the Current Population Survey. Washington, D.C.: U.S. Department of Commerce. Office of Federal Statistical Policy and Standards, Statistical Policy Working Paper 3.

Deming, W.E. (1982). Quality, Productivity, and Competitive Position. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Diskin, D. (1988). Automating Process Controls for Surveys and Censuses. U.S. Department of Commerce, Bureau of the Census, Statistical Methods Division.

Glover, C.S., Whitehouse, D.L., and Alexander, C.H. (1988). An Experiment with Process Control: Implementation, Results, and Recommendations. Washington, D.C., U.S. Census Bureau. Statistical Methods Division, Internal Report.

Rogot, E., Sorlie, P.D., Johnson, N.J., Glover, C.S., and Treasure, D.W. (1988). A Mortality Study of One Million Persons: First Data Book. Washington, D.C.: National Institute of Health. NIH Publication No. 88-2896.