Using Conversion Factors to Lower Transit Data Collection Costs

PETER G. FURTH AND BRIAN MCCOLLOM

Monitoring passenger use measures such as boardings, revenue, passenger-miles, and load can be an expensive task for transit systems. One way to reduce data collection costs is to measure only one of these items, the auxiliary item, and then apply multiplicative factors that are estimated from a small joint sample to its estimated mean to estimate the means of the other items. Statistical aspects of the conversion factor approach are presented, including sample size estimation and determination of the accuracy of both the conversion factors and the inferred estimates. An optimal sampling plan that minimizes the combined cost of estimating the conversion factors and estimating the mean of the auxiliary item is determined. Cost savings between 0 and 75 percent of the cost of direct estimation are obtained for various situations.

Collection of passenger use data (e.g., boardings, peak load, revenue) at both the route and system level constitutes a significant expense for transit systems. With the exception of system-level revenue, few transit systems collect these data on every trip. For cost reasons, sampling is used at most systems. Different sampling techniques are available, entailing different measurement techniques, sample sizes, sampling plans, and costs. One especially suitable technique for estimating passenger use measures is conversion factors. Although the informal use of conversion factors is widespread throughout the transit industry, the research effort reported in this paper represents the first treatment, to the authors' knowledge, of statistical issues, such as sample size and accuracy, related to the use of conversion factors in transit.

Although sampling has been done in the transit industry for decades—in bus systems, the abolition of tickets made it necessary—the subject of statistical determination of sample size and accuracy for transit has scarcely been started. A data collection manual published in 1947 by the American Transit Association *(1)*, recommends sample sizes but does not suggest the resultant level of precision. A more recent report *(2)* describes current data collection techniques but says nothing about sample size or accuracy.

In 1979, UMTA began the Bus Transit Monitoring Study to improve the state of the art in transit data collection and service monitoring practices. As part of that study, the first *Bus Transit Monitoring Manual (3)* was completed in 1981. A later edition of this manual was published in 1985 under the title *Transit Data Collection Design Manual* (TDCDM) *(4)*. In these manuals, procedures are offered for determining accuracy and sample size. The later edition includes procedures for using conversion factors under certain conditions that are described later. This paper draws on the results of the Bus Transit Monitoring Study and describes the statistical theory of conversion factors and their application to transit more fully.

P. G. Furth, Department of Civil Engineering, Northeastern University, Boston, Mass. 02115. B. McCollom, Comsis Corp., 2000 N. 15th St., Suite 507, Arlington, Va. 22201.

A conversion factor expresses a relationship between two variables: Y, called the inferred item, and X, the auxiliary item. The conversion factor is estimated from a paired sample of X and Y. The following notation is used in this paper:

N = population size,
Y_i = value of Y for population element i,
$Y_T = \sum_{i=1}^{N} Y_i$ = population total of Y,
$\bar{Y} = Y_T/N$ = population mean of Y,
n = size of paired sample,
y_i = value of Y for sample observation i,
$y_t = \sum_{i=1}^{n} y_i$ = sample total of Y,
$\bar{y} = y_T/n$ = sample mean of Y,
$S_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$ = variance of Y,
$v_y = \frac{S_y}{\bar{Y}}$ = coefficient of variation (COV) of Y,
and
$r_{xy} = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})(X_i - \bar{X})}{(N-1)S_y S_x}$ = correlation coefficient of X and Y.

Corresponding terms for X are similarly defined.

The conversion factor R is the ratio of population means. Its estimator $\hat{R}$ is the ratio of sample means. Thus,

$$R = \bar{Y}/\bar{X} = Y_T/X_T \quad (1)$$

$$\hat{R} = \bar{y}/\bar{x} = y_T/x_T \quad (2)$$

If the mean ($\bar{X}$) or total (X_T) of the auxiliary item is known, the mean or total of the inferred item can be estimated using the estimated conversion factor:

$$\hat{\bar{Y}} = \hat{R}\,\bar{X} \quad (3)$$

$$\hat{Y}_T = \hat{R}X_T \quad (4)$$

where $\hat{\bar{Y}}$ and $\hat{Y}_T$ are estimates of $\bar{Y}$ and Y_T. If $\bar{X}$ and X_T are not known but are estimates, the following estimates are used:

$$\hat{\bar{Y}} = \hat{R}\bar{x} \tag{5}$$

$$\hat{\bar{Y}}_T = \hat{R}\, N\bar{x} \tag{6}$$

Examples of conversion factors in the transit industry abound. Average fare is used to estimate route revenue from route boardings or vice versa. Average trip length is the conversion factor (average passenger-miles divided by average boardings) used to estimate passenger-miles from boardings data. Peak-load counts could be used to estimate route boardings. Evidence of the value of using ratio estimates is found in a recent study *(5)* of San Diego bus routes in which the ratio of maximum load to total boardings on a route showed far less variation over the day than the absolute values of either measure.

Because of the close relationship between such passenger use measures, a relatively small sample size is typically required for estimating conversion ratios. Armed with a set of ratios based on a single auxiliary item X, a transit system may be able to estimate all passenger use items by directly collecting data only on the variable X. This approach can be especially cost-effective when data on X can be collected much more cheaply than other passenger use measures.

Basic statistical theory is used in this paper and is therefore presented without proof. The theory can be found in several statistical texts; for example, Cochran *(6)*.

BIAS AND VARIANCE OF A CONVERSION FACTOR

The true model can be assumed as

$$Y_i = RX_i + \varepsilon_i \tag{7}$$

where ε_i is independently distributed with mean 0. Then $\hat{R}$ is a slightly biased estimator of R. Following Cochran *(6)*, this bias is of the order of $1/n$ and so is negligible for large samples. For sample sizes in the range of 10 to 30, terms of order $1/n^3$ and higher may be neglected, and so the relative mean squared error (mse) is approximately the expectation

$$\text{mse}(\hat{R}) = \text{E}\left[\frac{\hat{R} - R}{R}\right]^2$$

$$\cong v_R^2\left[1 + \frac{3v_x^2}{n} + \frac{6v_x^2}{n}\left(\frac{r_{xy}^2 v_y^2 + v_x^2 - 2r_{xy}v_x v_y}{v_y^2 + v_x^2 - 2r_{xy}v_x v_y}\right)\right] \tag{8}$$

where

$$v_R^2 = \frac{v_x^2 + v_y^2 - 2r_{xy}v_x v_y}{n} \tag{9}$$

is a first-order approximation of the relative variance (squared COV) of $\hat{R}$, derived from

$$v_R^2 \cong v_Y^2 \cong \frac{1}{n\bar{x}^2R^2}\,\frac{\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2}{n-1} \tag{10}$$

For many applications, X and Y have approximately equal coefficients of variation, so that Equation 7 reduces to

$$\text{mse}(\hat{R}) \cong v_R^2\left[1 + \frac{v_x^2}{n}(6 - 3r_{xy})\right] \tag{11}$$

By using values of $v_x = 0.6$ and $r_{xy} = 0.85$, worse than average values for route level applications observed in the UMTA Bus Transit Monitoring Study, v_R^2 underestimates the mse by a factor of about $(1 + 1.2/n)$.

The relative bias of $\hat{R}$, $b/\hat{R}$, where b = bias, can be expressed as

$$\frac{b}{R} = \left[\text{mse}\,(\hat{R}) - v_R^2\right]^{1/2} \tag{12}$$

When the assumed values of v_x and r_{xy} are used, the relative bias is approximately $0.36/n$. To keep the relative bias small, then, sample sizes below 10 should not be used.

Another problem with small samples is the fact that the approximations used become less exact because of nonnormality and the greater importance of neglected higher-order terms. Cochran *(6)* reports, based on empirical studies, that v_R^2 underestimates the true value by about 20 percent when the sample size is 10, and by about 6 percent when the sample size is 30. Accordingly multiplying the right-hand side of Equation 9 by the factor $n/(n-1.7)$ yields

$$v_R^2 = \frac{1}{n - 1.7}(v_x^2 + v_y^2 - 2r_{xy}v_x v_y) \tag{13}$$

ACCURACY AND SAMPLE SIZE

Accuracy is defined by two components, a relative tolerance and a confidence level. For example, the federal Section 15 requirements call for estimates of systemwide boardings and passenger-miles to within ±10 percent at a 95 percent confidence level. If an estimate $\hat{\theta}$ is normally distributed, then

$$d = z_c v(\hat{\theta}) \tag{14}$$

where

- d = tolerance (e.g., $d = 0.2$ means ±20 percent tolerance);
- c = specified confidence level in percent (e.g., $c = 95$ means 95 percent confidence level);
- z_c = $(c + 100)/2$ percentile value of the standard normal distribution (e.g., $z_{95} = 1.96 \cong 2$); and
- $v(\hat{\theta})$ = coefficient of variation of $\hat{\theta}$.

Although ratio estimates are not normally distributed, they are close enough to normal for practical purposes. Extreme nonnormality is avoided by prohibiting small sample sizes.

Because $v(\hat{\theta})$ depends on the sample size used to calculate $\hat{\theta}$, the sample size required to achieve a specified accuracy level can be determined from Equation 14. It is the sample size for which

$$v(\hat{\theta}) = d/z_c \tag{15}$$

It is assumed that N is known, so that the expansion from a mean to the population total does not alter the COV of an estimate, making the sample size formulas for means and totals identical. For simplicity, our exposition will deal with estimation of the population mean of Y.

Three different cases will be covered: when $\overline{X}$ the mean of the auxiliary item is known; when $\overline{X}$ is unknown and estimated by using data that are also used to estimate the conversion ratio; and when $\overline{X}$ is unknown and is estimated independent of the data used to estimate the conversion ratio.

INFERRING WITH A KNOWN AUXILIARY VARIABLE

The simplest case of using a conversion factor is when the total or mean of the auxiliary variable is known. An example is estimating systemwide boardings or passenger-miles from systemwide farebox revenue, which is universally measured for accounting purposes. Another transportation example might be estimating regional vehicle-miles of travel (VMT) from total regional gasoline sales.

When $\overline{X}$ is known, estimation of $\hat{\overline{Y}}$ follows Equation 3, and therefore

$$v^2(\hat{\overline{Y}}) = v^2(\hat{R}) = \frac{1}{n - 1.7}(v_x^2 + v_y^2 - 2r_{xy}v_xv_y) \tag{16}$$

Necessary sample size, applying Equation 15, is

$$n = \frac{z_c^2}{d^2}(v_x^2 + v_y^2 - 2r_{xy}v_xv_y) + 1.7 \tag{17}$$

The results of evaluating Equation 17 are always rounded upward to the next integer, with the smallest value for n of 10.

This approach was used to analyze the potential of using a systemwide revenue-to-boardings conversion in the collection of Section 15 data for the Port Authority of Allegheny County (PAT), Pittsburgh's regional transit system. The population was all the vehicle trips in a year including all routes; Y_i = boardings on Trip i and X_i = revenue on Trip i. Analysis of data from PAT yielded the following estimates:

$$v_x = 0.786, \qquad v_y = 0.587, \qquad r_{xy} = 0.647$$

When these parameters were used to obtain systemwide boardings with the specified ±10 percent tolerance at the 95 percent confidence level, n = 149.

This result indicates that, in the case of PAT, 149 trips must be randomly sampled over the year, with trip level boardings and revenue observed on each trip. This procedure might be done by using on-board checkers to count boarding passengers and to take farebox readings at the start and end of each trip.

An estimate of total passenger-miles at the same level of accuracy is also required in Section 15. An analysis was made of the potential of using a conversion factor of systemwide revenue (X) to passenger-miles (Y) to estimate total passenger-miles. An analysis of sample PAT data yielded the estimates v_y = 0.816 and r_{xy} = 0.739, resulting in a required sample size of n = 129. Data on both revenue and passenger-miles can be collected by having an on-board checker record ons and offs at each stop, from which passenger-miles may be derived directly, as well as the beginning and ending farebox reading. If on-board checks are used for estimating conversion factors for both passenger-miles and boardings, PAT could meet its Section 15 requirements for service consumed by performing annually 149 (the larger of the two n-values) on-board checks chosen at random, or about 3 per week.

In contrast, if boardings or passenger-miles are estimated directly without conversion for a sample of n trips, the COV of the sample mean would be $v_y/n^{1/2}$. Applying Equation 15, the necessary sample size would be

$$n = \frac{z_c^2}{d^2}v_y^2 \tag{18}$$

For the PAT example, Equation 18 yields n = 235 for boardings and n = 326 for passenger-miles. Assuming that data on both of these items can be measured concurrently with on-board checks, the larger sample size (n = 326) governs. In comparison with the 149 on-board checks needed under the conversion factor approach, use of revenue-based conversion factors saves, for this example, 56 percent of the data collection cost.

On the basis of the analysis of the PAT data described here and of a similar analysis of data from VIA (San Antonio's transit system), UMTA has authorized this revenue-based approach for meeting its Section 15 requirements for boardings and passenger-miles *(7)*. The approved sampling plan calls for 208 observations per year, in contrast with the 600 or so observations required in the previously authorized sampling plan that does not involve conversion factors *(8)*.

Some transit systems are fortunate in that route level boardings are counted routinely on every trip. This makes boardings an ideal auxiliary variable from which to estimate other desired route level measures such as average revenue per trip, average peak load, and average boardings on a segment of the route. For example, average peak load for a certain route/direction/time period ($R/D/TP$) combination (an example time period might be 6:30 to 9:00 a.m.) might be a desired statistic. Then the population is all trips in that $R/D/TP$, X_i = boardings on Trip i, and Y_i = peak load on Trip i. The boardings-to-peak-load conversion factor is estimated from a sample of n joint observations as $\hat{R} = \bar{y}/\bar{x}$. A joint observation is made by positioning a checker at the peak-load point to count the number of passengers on board. If the checker records the bus and trip number, this count can be matched to the boarding count taken on that trip. Analysis of data from a large number of routes in Pittsburgh, San Francisco, and Minneapolis yielded typical values as follows:

$$v_x = 0.5, \qquad v_y = 0.4, \qquad r_{xy} = 0.94$$

For a desired accuracy of ±10 percent tolerance at the 95 percent confidence level, the necessary sample size applying Equation 17 is n = 16.

In contrast, if peak load were estimated directly, the required number of observations of peak load, from Equation 18, would be $n = 64$. This four-fold difference illustrates the potential savings of using a known measure (e.g. boardings) that is being collected routinely to estimate other passenger-use measures by using conversion factors.

SAMPLING FOR THE CONVERSION FACTOR AND THE AUXILIARY ITEM IN THE SAME TIME FRAME

When there is no perfectly known auxiliary item, the mean of Y is estimated by using Equation 5, and both the conversion factor and the mean of the auxiliary item must be estimated by sampling. The case in which sampling for both items occurs in the same time frame is discussed in this section.

A random sample of n' observations of X are made. On a randomly selected subset of n of the observations, Y is observed as well. The following notation is used in this discussion:

$\bar{y}_{n'}, \bar{x}_n$ = sample means of Y and X from the subset of size n; and

$\bar{x}_{n'}$ = sample mean of X from the entire set of n' observations.

For example, if X is peak load for a particular $R/D/TP$, and Y is boardings for that $R/D/TP$, both X and Y might be observed by using on-board checks on n trips, whereas point checks are used to measure peak load on $n' - n$ additional trips. From Equation 2, the estimate of the peak load-to-boardings conversion factor is $\hat{R} = \bar{y}_n/\bar{x}_n$. The estimate of $\bar{X}$ is $\bar{x}_{n'}$. The estimate of $\bar{y}$ is

$$\hat{\bar{y}} = \bar{x}_{n'} \frac{\bar{y}_n}{\bar{x}_n} \tag{19}$$

Its squared COV, following Cochran and using the small-sample adjustment, is

$$v^2(\hat{\bar{Y}}) = \frac{v_x^2 + v_y^2 - 2r_{xy}v_xv_y}{n - 1.7} + \frac{2r_{xy}v_xv_y - v_x^2}{n'} \tag{20}$$

In determining the necessary sample size to achieve a desired accuracy, there is a trade-off between n and n'. The optimal choice depends on their relative costs. If the unit sampling cost for a paired observation is c, and the cost for a separate observation of X is c', the total cost is $cn + c'(n' - n)$.

If the following substitutions are made,

$$c_1 = c - c' \tag{21}$$

$$c_2 = c' \tag{22}$$

$$k_1 = v_x^2 + v_y^2 - 2r_{xy}v_xv_y \tag{23}$$

$$k_2 = 2r_{xy}v_xv_y - v_x^2 \tag{24}$$

the optimization problem is the following:

$$\text{Minimize } c_1n + c_2n' \tag{25}$$

subject to

$$k_1/(n - 1.7) + k_2/n' \le d^2/z^2 \tag{26}$$

$$n \ge 10 \tag{27}$$

and

$$n' > n. \tag{28}$$

The solution, ignoring Equation 28, is

$$n = 1.7 + \frac{z^2}{d^2}k_1\left[1 + \left(\frac{k_2c_2}{k_1c_1}\right)^{1/2}\right] \tag{29}$$

$$n' = k_2\left(\frac{d^2}{z^2} - \frac{k_1}{n - 1.7}\right)^{-1} \tag{30}$$

Results of Equation 29 are again to be rounded upwards, with the smallest allowed value for n of 10.

If $n' \le n$, Equation 28 is violated, implying that the direct estimation of $\bar{Y}$ without using a conversion factor will be less costly. Even if $n' > n$, there still may be cases where direct estimation of $\bar{Y}$ is the less costly alternative. Therefore the direct and indirect approaches must be compared.

An example of this approach can be shown using the data from the last example. The problem is to estimate mean boardings for a particular $R/D/TP$ to a tolerance of ±20 percent at a 95 percent confidence level through the use of a conversion factor with peak load as the auxiliary variate. The COV and correlation estimates are the same as for the last example, except that load is now the independent variable X and mean boardings is the dependent variable Y. The correlation coefficient is unchanged, because of the identity $r_{xy} = r_{yx}$. Joint observations of peak load and boardings are made using an on-board checker. If this checker will be collecting useful data in both directions, then the time required for one joint observation is $t/2$, where t = cycle time for the route. Observations of peak load alone are made with checkers stationed at the peak-load point. If the headway is h, one observation in each direction can usually be made in an interval h, so the time spent per observation is $h/2$. If $t = 60$ min and $h = 10$ min, then

$c_2 = h/2 = 5$ min

$c_1 = (t - h)/2 = 25$ min

$k_1 = 0.03$

$k_2 = 0.22$

$(z/d)^2 = 100$

$n = 8.33$, which is increased to 10,

$n' = 34$

The cost is $[cn + c'(n' - n)]/60 = 7$ checker-hr. For comparison, boardings could be estimated directly using on-board checkers. The unit sampling cost is $t/2 = 30$ min. The necessary sample size is, from Equation 18, $n = 25$. The resulting cost is 12.5 checker-hr. These results are summarized as follows:

	Using Conversion Factor (checker-hr)	*Direct Estimation (checker-hr)*
On-board checks	10	25
Point checks	24	0
Total	7	12.5

The conversion factor approach is not always the least costly alternative. Because a less strict accuracy for route boardings of ±30 percent at a 90 percent confidence level (z =1.65) is recommended in the TDCDM, it is illustrative to look at the results for this case. Applying Equation 29 yields n = 3.6, which must be rounded up to n = 10 on-board checks. Applying Equation 30 yields n' = 7.5, which is less than n, indicating that no point checks should be taken and that direct estimation will be less costly. When Equation 18 is applied, it is found that direct estimation calls for only n = 8 ride checks in this case. This example illustrates how the requirement of having at least 10 joint observations makes the conversion factor approach less desirable when only small sample sizes are needed.

SAMPLING FOR THE CONVERSION FACTOR AND THE AUXILIARY ITEM INDEPENDENTLY

In some cases the conversion factor and the mean of the auxiliary item can be estimated from independent samples. For example, the peak load-to-boardings conversion factor might be estimated in the fall and used with an estimate of peak load in the spring to estimate spring mean boardings. In this case, the conversion factor is estimated from a set of n joint observations of X and Y with unit sampling cost c, yielding the conversion factor estimate $\hat{R} = \bar{y}_n/\bar{x}_n$ as before. An independent sample of n' observations of X with unit sampling cost c' yields the estimate of $\bar{x}_{n'}$. The product of $\hat{R}$ and $\bar{x}_{n'}$ is an estimate of $\bar{Y}$, whose squared COV, following Cochran *(6)*, is approximately

$$v^2(\hat{\bar{Y}}) = \frac{v_y^2 + v_x^2 - 2r_{xy}v_xv_y}{n - 1.7} + \frac{v_x^2}{n'} \tag{31}$$

If a conversion factor has already been estimated by using a sample size n, it follows that the required size of the sample of the auxiliary item is

$$n' = v_x^2\left(\frac{d^2}{z^2} - \frac{v_x^2 + v_y^2 - 2r_{xy}v_xv_y}{n - 1.7}\right)^{-1} \tag{32}$$

provided that the denominator is greater than zero. If the denominator is negative, the desired accuracy level is unattainable regardless of n', because there is too much uncertainty in the estimate of the conversion factor.

A program for route level monitoring in which conversion factors are estimated during a baseline phase and are assumed to remain stable during a monitoring phase of a few years is recommended in the TDCDM. During the monitoring phase, $\bar{X}$ is estimated periodically (e.g., every quarter) and is multiplied by the conversion factor to yield periodic estimates of $\bar{Y}$.

Cost minimization should be considered in the selection of the sample sizes for the auxiliary item that will be taken during the monitoring phase and for the one conversion factor sample taken in the baseline phase. If f = number of times a sample of $\bar{X}$ will be measured during the monitoring phase, then the problem of minimizing the total data collection cost of both baseline and monitoring phases is determined from the following problem:

$$\text{minimize } cn + fc'n' \tag{33}$$

subject to

$$\frac{v_y^2 + v_x^2 - 2r_{xy}v_xv_y}{n - 1.7} + \frac{v_x^2}{n'} \le \frac{d^2}{z^2} \tag{34}$$

$$n \ge 10 \tag{35}$$

This problem is of the same form as Equations 25–27. Therefore, the solutions for n (Equation 29) and n' (Equation 30) can be used by making the following substitutions:

$$c_1 = c \tag{36}$$

$$c_2 = fc' \tag{37}$$

$$k_1 = v_x^2 + v_y^2 - 2r_{xy}v_xv_y \tag{38}$$

$$k_2 = v_x^2 \tag{39}$$

This approach can be applied to the previously used example, in which the problem is to estimate mean boardings Y for a particular $R/D/TD$ from mean peak load X. The desired tolerance for mean boardings is ±20 percent at the 95 percent confidence level. If the monitoring phase lasts 3 years and a quarterly estimate of $\bar{Y}$ is needed, then f = 12. Using the same values for v_x, v_y, and r_{xy} as before, n and n' are estimated as follows:

c_1 = 30 min

c_2 = 60 min

k_1 = 0.03

$v_x^2 = k_2$ = 0.25

$(z/d)^2$ = 100

n = 17

n' = 31

Using Equation 33, the cost for both the baseline and monitoring phases is estimated to be 39.5 checker-hr. In comparison, direct estimation of boardings, using ride checks, during the monitoring phase requires (as before) 25 observations for each estimate of Y, for a cost of 150 checker-hr. This cost is in addition to any cost of sampling in the baseline year.

If the same example is repeated with a ±30 percent tolerance at the 90 percent confidence level required, the results are

	Using Conversion Factors (checker-hr)	*Direct Estimation (checker-hr)*
n	10	8
n'	9	–
Total cost	14	48

In both cases, conversion factors offer a 70 to 75 percent reduction in data collection costs over direct estimation.

USING AN AUXILIARY ITEM TO INFER SEVERAL ITEMS

In the transit context, a single auxiliary item may be used with different conversion factors to yield estimates of several items, such as route boardings, segment boardings, peak load, and revenue. For each estimated item there is a different set of values for v_y, r_{xy}, and d. What, then, is the least-cost sampling plan?

Expanding the optimization framework is straightforward if all of the conversion factors are estimated using the same data collection technique, typically ride checks. If the auxiliary item is known without sampling, the estimated variable that demands the largest paired sample size will control the sample size.

When the auxiliary item as well as several conversion factors must be estimated by sampling, the general solution is cumbersome to apply because several optimality conditions must be checked. One simple approach is to do a one-dimensional search over n, the number of ride checks. The lower bound for n is 10. If the conversion factor sample is independent of the auxiliary item sample, each inferred item will supply another lower bound for n, which is the integer greater than the value at which the denominator of Equation 32 equals 0. Then, for a given value of n, the required n' for each inferred item is given by Equation 30 or 32. The highest n' governs. With the given n and corresponding governing n', the cost can easily be calculated. Searching over n will reveal a value that minimizes overall cost. Although not proven, the cost as a function of n is almost certain to be convex. This implies that a local optimum will be a global optimum.

CONCLUSIONS

In this paper, formulas for sample size and accuracy determination are presented with the use of conversion factors in three situations: (a) when the mean $\overline{X}$ of the auxiliary item is known; (b) when it is estimated using the conversion factor sample plus a supplemental sample of X; and (c) when $\overline{X}$ is estimated independently of the conversion factor sample. In the latter cases, the trade-off between the costs of the conversion factor sample and of the sample for X is explored, and cost-minimizing formulas are presented. Examples are presented that demonstrate how using conversion factors can dramatically reduce data collection costs in comparison with direct estimation.

ACKNOWLEDGMENTS

Parts of this research were funded by UMTA. The authors are indebted to Nigel Wilson of the Massachusetts Institute of Technology, who initiated research on this topic.

REFERENCES

1. W. S. Rainville, Jr. *Traffic Checking and Schedule Preparation.* American Transit Association, 1947. (Reprinted by U.S. Department of Transportation, as Report DOT–I–88–23, 1982).
2. *Review of Transit Data Collection Techniques.* Report DOT–I–85–26, UMTA, U.S. Department of Transportation, 1985.
3. J. P. Attanucci, I. Burns, and N. H. Wilson. *Bus Transit Monitoring Manual.* Report DOT–I–81–30, U.S. Department of Transportation, 1981.
4. J. H. Banks. Analysis of Maximum Load Data for an Urban Bus System. In *Transportation Research Record 1011,* TRB, National Research Council, Washington, D.C., 1985, pp. 1–8.
5. P. G. Furth, J. P. Attanucci, I. Burns, and N. H. Wilson. *Transit Data Collection Design Manual.* U.S. Department of Transportation, 1985.
6. W. G. Cochran, *Sampling Techniques.* 3rd edition. John Wiley and Sons, Inc., New York, 1977.
7. *Revenue Based Sampling Procedures for Obtaining Fixed Route Bus Operating Data Required Under the Section 15 Reporting System.* Circular UMTA–C–2710.4, UMTA, U.S. Department of Transportation, 1985.
8. *Sampling Procedures for Obtaining Fixed Route Bus Operating Data Required Under the Section 15 Reporting System.* Circular UMTA–C–2710.1, UMTA, U.S. Department of Transportation, 1978.

Publication of this paper sponsored by Committee on Transit Management and Performance.