Chapter 4

A Method for Calibrating False-Match Rates in Record Linkage*

Thomas R. Belin, UCLA and Donald B. Rubin, Harvard University

Specifying a record-linkage procedure requires both (1) a method for measuring closeness of agreement between records, typically a scalar weight, and (2) a rule for deciding when to classify records as matches or nonmatches based on the weights. Here we outline a general strategy for the second problem, that is, for accurately estimating false-match rates for each possible cutoff weight. The strategy uses a model where the distribution of observed weights are viewed as a mixture of weights for true matches and weights for false matches. An EM algorithm for fitting mixtures of transformed-normal distributions is used to find posterior modes; associated posterior variability is due to uncertainty about specific normalizing transformations as well as uncertainty in the parameters of the mixture model, the latter being calculated using the SEM algorithm. This mixture-model calibration method is shown to perform well in an applied setting with census data. Further, a simulation experiment reveals that, across a wide variety of settings not satisfying the model's assumptions, the procedure is slightly conservative on average in the sense of overstating false-match rates, and the one-sided confidence coverage (i.e., the proportion of times that these interval estimates cover or overstate the actual false-match rate) is very close to the nominal rate.

KEY WORDS: Box–Cox transformation; Candidate matched pairs; EM algorithm; Mixture model; SEM algorithm; Weights.

1. AN OVERVIEW OF RECORD LINKAGE AND THE PROBLEM OF CALIBRATING FALSE-MATCH RATES

1.1 General Description of Record Linkage

Record linkage (or computer matching, or exact matching) refers to the use of an algorithmic technique to identify records from different data bases that correspond to the same individual. Record-linkage techniques are used in a variety of settings; the current work was formulated and first applied in the context of record linkage between the census and a large-scale postenumeration survey (the PES), which comprises the first step of an extensive matching operation conducted to evaluate census coverage for subgroups of the population (Hogan 1992). The goal of this first step is to declare as many records as possible "matched" without an excessive rate of error, thereby avoiding the cost of the resulting manual processing for all records not declared "matched."

Specifying a record-linkage procedure requires both a method for measuring closeness of agreement between records and a rule using this measure for deciding when to classify records as matches. Much attention has been paid in the record-linkage literature to the problem of assigning "weights" to individual fields of information in a multivariate record and obtaining a "composite weight" that summarizes the closeness of agreement between two records (see, for example, Copas and Hilton 1990; Fellegi and Sunter 1969; Newcombe 1988; and Newcombe, Kennedy, Axford, and James 1959). Somewhat less attention has been paid to the problem of deciding when to classify records as matches, although various approaches have been offered by Tepping (1968), Fellegi and Sunter (1969), Rogot, Sorlie, and Johnson (1986), and Newcombe (1988). Our work focuses on the second problem by providing a predicted probability of match for two records, with associated standard error, as a function of the composite weight.

The context of our problem, computer matching of census records, is typical of record linkage. After data collection, preprocessing of data, and determination of weights, the next step is the assignment of candidate matched pairs where each pair of records consists of the best potential match for each other from the respective data bases (cf. "hits" in Rogot, Sorlie, and Johnson 1986; "pairs" in Winkler 1989; "assigned pairs" in Jaro 1989). According to specified rules, a scalar weight is assigned to each candidate pair, thereby ordering the pairs. The final step of the record linkage procedure is viewed as a decision problem where three actions are possible for each candidate matched pair: declare the two records matched, declare the records not matched, or send both records to be reviewed more closely (see, for example, Fellegi and Sunter 1969). In the motivating problem at the U.S. Census Bureau, a binary choice is made between the alternatives "declare matched" versus "send to followup," although the matching procedure attempts to draw distinctions within the latter group to make manual matching easier for follow-up clerks. In such a setting, a cutoff weight is needed above which records are declared matched; the false-match rate is then defined as the number of falsely matched pairs divided by the number of declared matched pairs. Particularly relevant for any such decision problem is an accurate method for assessing the probability that a candidate matched pair is a correct match as a function of its weight.

1.2 The Need for Better Methods of Classifying Records as Matches or Nonmatches

Belin (1989a, 1989b, 1990) studied various weighting procedures (including some suggested by theory, some used in practice, and some new simple ad hoc weighting schemes) in the census matching problem and reached three primary

* Thomas R. Belin is Assistant Professor, Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA, 90024. Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA, 02138. The authors would like to thank William Winkler of the U.S. Census Bureau for a variety of helpful discussions. The authors also gratefully acknowledge the support of Joint Statistical Agreements 89-07, 90-23, and 91-08 between the Census Bureau and Harvard University, which helped make this research possible. Much of this work was done while the first author was working for the Record Linkage Staff of the Census Bureau; the views expressed are those of the authors and do not necessarily reflect those of the Census Bureau.

© 1995 American Statistical Association
Journal of the American Statistical Association
June 1995, Vol. 90, No. 430, Theory and Methods
Reprinted with permission.

conclusions. First, different weighting procedures lead to comparable accuracy in computer matching. Second, as expected logically and from previous work (e.g., Newcombe 1988, p. 144), the false-match rate is very sensitive to the setting of a cutoff weight above which records will be declared matched. Third, and more surprising, current methods for estimating the false-match rate are extremely inaccurate, typically grossly optimistic.

To illustrate this third conclusion, Table 1 displays empirical findings from Belin (1990) with test-census data on the performance of the procedure of Fellegi and Sunter (1969), which relies on an assumption of independence of agreement across fields of information. That the Fellegi–Sunter procedure for estimating false-match rates does not work well (i.e., is poorly calibrated) may not be so surprising in this setting, because the census data being matched do not conform well to the model of mutual independence of agreement across the fields of information (see, for example, Kelley 1986 and Thibaudeau 1989). Other approaches to estimating false-match rates that rely on strong independence assumptions (e.g., Newcombe 1988) can be criticized on similar grounds.

Although the Fellegi–Sunter approach to setting cutoff weights was originally included in census/PES matching operations (Jaro 1989), in the recent past (including in the 1990 Census) the operational procedure for classifying record pairs as matches has been to have a human observer establish cutoff weights manually by "eyeballing" lists of pairs of records brought together as candidate matches. This manual approach is easily criticized, both because the error properties of the procedure are unknown and variable and because, when linkage is done in batches at different times or by different persons, inconsistent standards are apt to be applied across batches.

Another idea is to use external data to help solve this classification problem. For example, Rogot, Sorlie, and Johnson (1986) relied on extreme order statistics from pilot data to determine cutoffs between matches and nonmatches; but this technique can be criticized, because extreme order statistics may vary considerably from sample to sample, especially when sample sizes are not large. One other possibility, discussed by Tepping (1968), requires clerical review of samples from the output of a record-linkage procedure to provide feedback on error rates to refine the calibration procedure. Such feedback is obviously desirable, but in many applications, including the census/PES setting, it is impossible to provide it promptly enough.

Table 1. Performance of Fellegi–Sunter Cutoff Procedure on 1986 Los Angeles Test-Census Data

Acceptable false-match rate specified by user of matching program	Observed false-match rate among declared matched pairs
.05	.0627
.04	.0620
.03	.0620
.02	.0619
.01	.0497
10^{-3}	.0365
10^{-4}	.0224
10^{-5}	.0067
10^{-6}	.0067
10^{-7}	.0067

A more generally feasible strategy is to use the results of earlier record-linkage studies in which all candidate matched pairs have been carefully reviewed by clerks. This type of review is common practice in operations conducted by the Census Bureau. Each such training study provides a data set in which each candidate pair has its weight and an outcome, defined as true match or false match, and thus provides information for building a model to give probability of match as a function of weight.

1.3 A Proposed Solution to the Problem of Calibrating Error Rates

There are two distinct approaches to estimating the relationship between a dichotomous outcome, $Z_i = 1$ if match and $Z_i = 0$ if nonmatch, from a continuous predictor, the weight, W_i: the direct approach, typified by logistic regression, and the indirect approach, typified by discriminant analysis. In the direct approach, an iid model is of the form $f(Z_i \mid W_i, \nu) \times g(W_i \mid \zeta)$, where $g(W_i \mid \zeta)$, the marginal distribution of W_i, is left unspecified with ζ a priori independent of ν. In the indirect approach, the iid model is of the form $h(W_i \mid Z_i, \phi)[\lambda^{Z_i}(1-\lambda)^{(1-Z_i)}]$, where the first factor specifies, for example, a normal conditional distribution of W_i for $Z_i = 0$ and for $Z_i = 1$ with common variance but different means, and the second factor specifies the marginal probability of $Z_i = 1$, λ, which is a priori independent of ϕ. Under this approach, $P(Z_i \mid W_i)$ is found using Bayes's theorem from the other model specifications as a function of ϕ and λ. Many authors have discussed distinctions between the two approaches, including Halperin, Blackwelder, and Verter (1971), Mantel and Brown (1974), Efron (1975), and Dawid (1976).

In our setting, application of the direct approach would involve estimating $f(Z_i \mid W_i, \nu)$ in observed sites where determinations of clerks had established Z_i, and then applying the estimated value of ν to the current site with only W_i observed to estimate the probability of match for each candidate pair. If the previous sites differed only randomly from the current sites, or if the previous sites were a subsample of the current data selected on W_i, then this approach would be ideal. Also, if there were many previous sites and each could be described by relevant covariates, such as urban/rural and region of the country, then the direct approach could estimate the distribution of Z as a function of W and covariates and could use this for the current site. Limited experience of ours and of our colleagues at the Census Bureau, who investigated this possibility using 1990 Census data, has resulted in logistic regression being rejected as a method for estimating false-match rates in the census setting (W. E. Winkler 1993, personal communication).

But the indirect approach has distinct advantages when, as in our setting, there can be substantial differences among sites that are not easily modeled as a function of covariates and we have substantial information on the distribution of weights given true and false matches, $h(\cdot \mid \cdot)$. In particular,

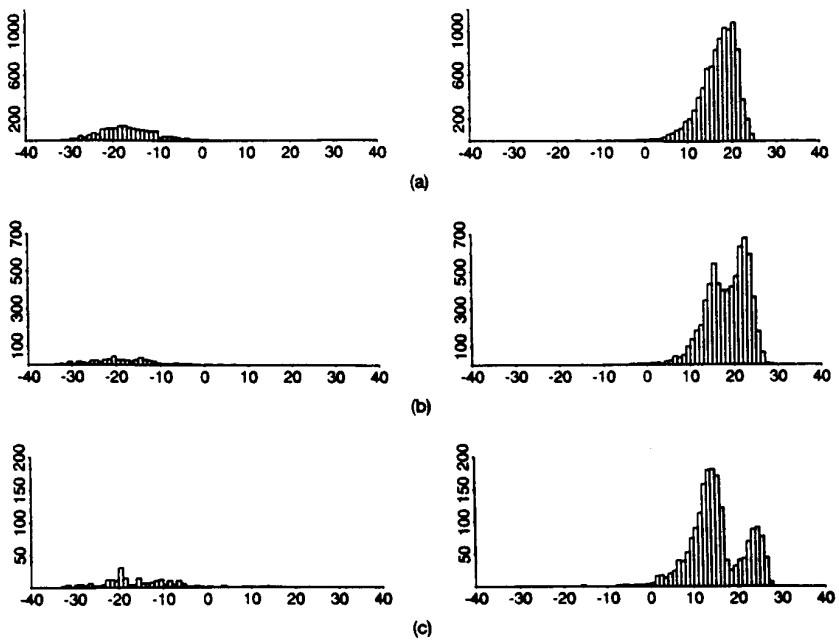

Figure 1. Histograms of Weights for True and False Matches by Site: (a) St. Louis; (b) Columbia; (c) Washington.

with the indirect approach, the observed marginal distribution of W_i in the current site is used to help estimate $P(Z_i \mid W_i)$ in this site, thereby allowing systematic site to site differences in $P(Z_i \mid W_i)$. In addition, there can be substantial gains in efficiency using the indirect approach when normality holds (Efron 1975), especially when $h(W_i \mid Z_i = 1, \phi)$ and $h(W_i \mid Z_i = 0, \phi)$ are well separated; (that is, when the number of standard deviations difference between their means is large).

Taking this idea one step further, suppose that previous validated data sets had shown that after a known transformation, the true-match weights were normally distributed, and that after a different known transformation, the false-match weights were normally distributed. Then, after inverting the transformations, $P(Z_i \mid W_i)$ could be estimated in the current site by fitting a normal mixture model, which would estimate the means and variances of the two normal components (i.e., ϕ) and the relative frequency of the two components (i.e., λ), and then applying Bayes's theorem. In this example, instead of assuming a common $P(Z_i \mid W_i)$ across all sites, only the normality after the fixed transformations would be assumed common across sites. If there were many sites with covariate descriptors, then (λ, ϕ) could be modeled as a function of these, for example, a linear model structure on the normal means.

To illustrate the application of our work, we use available test-census data consisting of records from three separate sites of the 1988 dress rehearsal Census and PES: St. Louis, Missouri, with 12,993 PES records; a region in East Central Missouri including Columbia, Missouri, with 7,855 PES records; and a rural area in eastern Washington state, with only 2,318 records. In each site records were reviewed by clerks, who made a final determination as to the actual match status of each record; for the purpose of our discussion, the clerks' determinations about the match status of record pairs are regarded as correct. The matching procedures used in the 1988 test Census have been documented by Brown et al. (1988), Jaro (1989), and Winkler (1989). Beyond differences in the sizes of the PES files, the types of street addresses in the areas offer considerably different amounts of information for matching purposes; for instance, rural route addresses, which were common in the Washington site but almost nonexistent in the St. Louis site, offer less information for matching than do most addresses commonly found in urban areas.

Figure 1 shows histograms of both true-match weights and false-match weights from each of the three sites. The bimodality in the true-match distribution for the Washington site appears to be due to some record pairs agreeing on address information and some not agreeing. This might generate concern, not so much for lack of fit in the center of the distribution as for lack of fit in the tails, which are essential to false-match rate estimation. Of course, it is not surprising that validated data dispel the assumption of normality for true-match weights and false-match weights. They do, however—at least at a coarse level in their apparent skewness—tend to support the idea of a similar nonnormal distributional shape for true-match weights across sites as well as a similar nonnormal distributional shape for false-match weights across sites. Moreover, although the locations of these distributions change from site to site, as do the relative frequencies of the true-match to the false-match components, the relative spread of the true to false components is similar across sites.

These observations lead us to formulate a transformed-normal mixture model for calibrating false-match rates in record-linkage settings. In this model, two power (or Box–Cox) transformations are used to normalize the false-match weights and the true-match weights, so that the observed raw weights in a current setting are viewed as a mixture of two transformed normal observations.

Mixture models have been used in a wide variety of statistical applications (see Titterington, Smith, and Makov 1985, pp. 16–21, for an extensive bibliography). Power transformations are also used widely in statistics, prominently in an effort to satisfy normal theory assumptions in regression settings (see, for example, Weisberg 1980, pp. 147–151). To our knowledge, neither of these techniques has been utilized in record linkage operations, nor have mixtures of transformed-normal distributions, with different transformations in the groups, appeared previously in the statistical literature, even though this extension is relatively straightforward. The most closely related effort to our own of which we are aware is that of Maclean, Morton, Elston, and Yee (1976), who used a common power transformation for different components of a mixture model, although their work focused on testing for the number of mixture components.

Section 2 describes the technology for fitting mixture models with components that are normally distributed after application of a power transformation, which provides the statistical basis for the proposed calibration method. This section also outlines the calibration procedure itself, including the calculation of standard errors for the predicted false-match rate. Section 3 demonstrates the performance of the method in the applied setting of matching the Census and PES, revealing it to be quite accurate. Section 4 summarizes a simulation experiment to gauge the performance of the calibration procedure in a range of hypothetical settings and this too supports the practical utility of the proposed calibration approach. Section 5 concludes the article with a brief discussion.

2. CALIBRATING FALSE-MATCH RATES IN RECORD LINKAGE USING TRANSFORMED-NORMAL MIXTURE MODELS

2.1 Strategy Based on Viewing Distribution of Weights as Mixture

We assume that a univariate composite weight has been calculated for each candidate pair in the record-linkage problem at hand, so that the distribution of observed weights is a mixture of the distribution of weights for true matches and the distribution of weights for false matches. We also assume the availability of at least one training sample in which match status (i.e., whether a pair of records is a true match or a false match) is known for all record pairs. In our applications, training samples come from other geographical locations previously studied. We implement and study the following strategy for calibrating the false-match rate in a current computer-matching problem:

1. Use the training sample to estimate "global" parameters, that is, the parameters of the transformations that normalize the true- and false-match weight distributions and the parameter that gives the ratio of variances between the two components on the transformed scale. The term "global" is used to indicate that these parameters are estimated by data from other sites and are assumed to be relatively constant from site to site, as opposed to "site-specific" parameters, which are assumed to vary from site to site and are estimated only by data from the current site.

2. Fix the values of the global parameters at the values estimated from the training sample and fit a mixture of transformed-normal distributions to the current site's weight data to obtain maximum likelihood estimates (MLE's) and associated standard errors of the component means, component variances, and mixing proportion. We use the EM algorithm (Dempster, Laird, and Rubin 1977) to obtain MLE's and the SEM algorithm (Meng and Rubin 1991) to obtain asymptotic standard errors.

3. For each possible cutoff level for weights, obtain a point estimate for the false-match rate based on the parameter estimates from the model and obtain an estimate of the standard error of the false-match rate. In calculating standard errors, we rely on a large-sample approximation that makes use of the estimated covariance matrix obtained from the SEM algorithm.

An appealing modification of this approach, which we later refer to as our "full strategy," reflects uncertainty in global parameters through giving them prior distributions. Then, rather than fixing the global parameters at their estimates from the training sample, we can effectively integrate over the uncertainty in the global parameters by modifying Step 2 to be:

2′. Draw values of the global parameters from their posterior distribution given training data, fix global parameters at their drawn values, and fit a mixture of transformed-normal distributions to the current weight data to obtain MLE's (and standard errors) of site-specific parameters;

and adding:

4. Repeat Steps 2′ and 3 a few or several times, obtaining false-match rate estimates and standard errors from each repetition, and combine the separate estimates and standard errors into a single point estimate and standard error that reflect uncertainty in the global parameters using the multiple imputation framework of Rubin (1987).

We now describe how to implement each of these steps.

2.2 Using a Training Sample to Estimate Global Parameters

Box and Cox (1964) offered two different parameterizations for the power family of transformations: one that ignores the scale of the observed data, and the other—which we will use—that scales the transformations by a function of the observed data so that the Jacobian is unity. We denote the family of transformations by

$$\psi(w_i; \gamma, \omega) = \frac{w_i^\gamma - 1}{\gamma\omega^{\gamma-1}} \quad \text{if } \gamma \neq 0,$$
$$= \omega \log(w_i) \quad \text{if } \gamma \neq 0, \qquad (1)$$

where ω is the geometric mean of the observations $w_1, \ldots, w_n$.

By "transformed-normal distribution," we mean that for some unknown values of γ and ω, the transformed observations $\psi(w_i; \gamma, \omega)$ $(i = 1, \ldots, n)$ are normally distributed. Although the sample geometric mean is determined by the data, we will soon turn to a setting involving a mixture of two components with different transformations to normality, in which even the sample geometric means of the two components are unknown; consequently, we treat ω as an unknown parameter, the population geometric mean.

When the transformations are not scaled by the geometric-mean factor, as Box and Cox (1964, p. 217) noted, "the general size and range of the transformed observations may depend strongly on $[\gamma]$." Of considerable interest in our setting is that when transformations are scaled, not only are the likelihoods for different values of γ directly comparable, at least asymptotically, but also, by implication, so are the residual sums of squares on the transformed scales for different values of γ. In other words, scaling the transformations by $\omega^{\gamma-1}$ has the effect asymptotically of unconfounding the estimated variance on the transformed scale from the estimated power parameter. This result is important in the context of fitting mixtures of transformed-normal distributions when putting constraints on component variances in the fitting of the mixture model; by using scaled transformations, we can constrain the variance ratio without reference to the specific power transformation that has been applied to the data.

Box and Cox (1964) also considered an unknown location parameter in the transformation, which may be needed because power transformations are defined only for positive random variables. Because the weights that arise from record-linkage procedures are often allowed to be negative, this issue is relevant in our application. Nevertheless, Belin (1991) reported acceptable results using an ad hoc linear transformation of record-linkage weights to a range from 1 to 1,000. Although this ad hoc shift and rescaling is assumed to be present, we suppress the parameters of this transformation in the notation.

In the next section we outline in detail a transformed-normal mixture model for record-linkage weights. Fitting this model requires separate estimates of γ and ω for the true-match and false-match distributions observed in the training data, as well as an estimate of the ratio of variances on the transformed scale. The γ's can, as usual, be estimated by performing a grid search of the likelihoods or of the respective posterior densities. A modal estimate of the variance ratio can be obtained as a by-product of the estimation of the γ's. We also obtain approximate large-sample variances by calculating for each parameter a second difference as numerical approximation to the second derivative of the log-likelihood in the neighborhood of the maximum (Belin 1991). In our work we have simply fixed the ω's at their sample values, which appeared to be adequate based on the overall success of the methodology on both real and simulated data; were it necessary to obtain a better fit to the data, this approach could be modified.

2.3 Fitting Transformed Normal Mixtures with Fixed Global Parameters

2.3.1 Background on Fitting Normal Mixtures Without Transformations. Suppose that f_1 and f_2 are densities that depend on an unknown parameter ϕ, and that the density f is a mixture of f_1 and f_2, i.e., $f(X|\phi, \lambda) = \lambda f_1(X|\phi) + (1-\lambda)f_2(X|\phi)$ for some λ between 0 and 1. Given an iid sample $(X_1, X_2, \ldots, X_n)$ from $f(X|\phi, \lambda)$, the likelihood of $\theta = (\phi, \lambda)$ can then be written as

$$L(\theta|X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i|\theta)$$
$$= \prod_{i=1}^{n} [\lambda f_1(X_i|\phi) + (1-\lambda)f_2(X_i|\phi)].$$

Following the work of many authors (e.g., Aitkin and Rubin 1985; Dempster et al. 1977; Little and Rubin 1987; Orchard and Woodbury 1972; Titterington et al. 1985), we formulate the mixture model in terms of unobserved indicators of component membership Z_i, $i = 1, \ldots, n$, where $Z_i = 1$ if X_i comes from component 1 and $Z_i = 0$ if X_i comes from component 2. The mixture model can then be expressed as a hierarchical model,

$$X_i|\{Z_i = 1\}, \theta \overset{\text{iid}}{\sim} f_1(\cdot|\phi)$$
$$X_i|\{Z_i = 0\}, \theta \overset{\text{iid}}{\sim} f_2(\cdot|\phi)$$
$$Z_i|\theta \overset{\text{iid}}{\sim} \text{Bernoulli}(\lambda).$$

The "complete-data" likelihood, which assumes that the "missing data" $Z_1, \ldots, Z_n$ are observed, can be written as

$$L(\phi, \lambda|X_1, \ldots, X_n; Z_1, \ldots, Z_n)$$
$$= \prod_{i=1}^{n} [\lambda f_1(X_i|\phi)]^{Z_i}[(1-\lambda)f_2(X_i|\phi)]^{1-Z_i}.$$

Viewing the indicators for component membership as missing data motivates the use of the EM algorithm to obtain MLE's of (ϕ, λ). The E step involves finding the expected value of the Z_i's given the data and current parameter estimates $\phi^{(t)}$ and $\lambda^{(t)}$, where t indexes the current iteration. This is computationally straightforward both because the iid structure of the model implies that Z_i is conditionally independent of the rest of the data given X_i and because the Z_i's are indicator variables, so the expectation of Z_i is simply the posterior probability that Z_i equals 1. Using Bayes's theorem, the E step at the $(t+1)$st iteration thus involves calculating

$$Z_i^{(t+1)} = E(Z_i|X_1, \ldots, X_n; \phi^{(t)}, \lambda^{(t)})$$
$$= \frac{\lambda^{(t)} f_1(X_i|\phi^{(t)})}{\lambda^{(t)} f_1(X_i|\phi^{(t)}) + (1-\lambda^{(t)})f_2(X_i|\phi^{(t)})} \quad (2)$$

for $i = 1, \ldots, n$.

The M step involves solving for MLE's of θ in the "complete-data" problem. In the case where f_1 corresponds to the $N(\mu_1, \sigma_1^2)$ distribution and f_2 corresponds to the $N(\mu_2, \sigma_2^2)$ distribution, so that $\phi = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, the M step at

iteration $(t+1)$ involves calculating

$$\mu_1^{(t+1)} = \frac{\sum_{i=1}^n Z_i^{(t+1)} X_i}{\sum_{i=1}^n Z_i^{(t+1)}}$$

$$\mu_2^{(t+1)} = \frac{\sum_{i=1}^n (1 - Z_i^{(t+1)}) X_i}{\sum_{i=1}^n (1 - Z_i^{(t+1)})} \tag{3}$$

and

$$\sigma_1^{2(t+1)} = \frac{\sum_{i=1}^n Z_i^{(t+1)} (X_i - \mu_1^{(t+1)})^2}{\sum_{i=1}^n Z_i^{(t+1)}}$$

$$\sigma_2^{2(t+1)} = \frac{\sum_{i=1}^n (1 - Z_i^{(t+1)})(X_i - \mu_2^{(t+1)})^2}{\sum_{i=1}^n (1 - Z_i^{(t+1)})}. \tag{4}$$

The updated value of λ at the $(t+1)$st iteration is given by

$$\lambda^{(t+1)} = \frac{1}{n} \sum_{i=1}^n Z_i^{(t+1)}, \tag{5}$$

which holds no matter what the form of the component densities may be. Instabilities can arise in maximum likelihood estimation for normally distributed components with distinct variances, because the likelihood is unbounded at the boundary of the parameter space where either $\sigma_i^2 = 0$. Unless the starting values for EM are near a local maximum of the likelihood, EM can drift toward the boundary where the resulting fitted model suggests that one component consists of any single observation (with zero variance) and that the other component consists of the remaining observations (Aitkin and Rubin 1985).

When a constraint is placed on the variances of the two components, EM will typically converge to an MLE in the interior of the parameter space. Accordingly, a common approach in this setting is to find a sensible constraint on the variance ratio between the two components or to develop an informative prior distribution for the variance ratio. When the variance ratio $V = \sigma_1^2/\sigma_2^2$ is assumed fixed, the E step proceeds exactly as in (2) and the M step for $\mu_1^{(t+1)}$ and $\mu_2^{(t+1)}$ is given by (3); the M step for the scale parameters with fixed V is

$$\sigma_1^{2(t+1)} = \frac{1}{n}\left[\sum_{i=1}^n Z_i^{(t+1)}(X_i - \mu_1^{(t+1)})^2 + V(1 - Z_i^{(t+1)}) \times (X_i - \mu_2^{(t+1)})^2\right], \qquad \sigma_2^{2(t+1)} = \frac{\sigma_1^{2(t+1)}}{V}. \tag{6}$$

2.3.2 Modifications to Normal Mixtures for Distinct Transformations of the Two Components. We now describe EM algorithms for obtaining MLE's of parameters in mixtures of transformed-normal distributions, where there are distinct transformations of each component. Throughout the discussion, we will assume that there are exactly two components; fitting mixtures of more than two components involves straightforward extensions of the arguments that follow (Aitkin and Rubin 1985).

We will also assume that the transformations are fixed; that is, we assume that the power parameters (the two γ_i's) and the "geometric-mean" parameters (the two ω_i's) are known in advance and are not to be estimated from the data. We can write the model for a mixture of transformed-normal components as follows:

$$X_i \mid \theta, Z_i = 1 \sim \text{Transformed-}N(\mu_1, \sigma_1^2, \gamma_1, \omega_1),$$

$$X_i \mid \theta, Z_i = 0 \sim \text{Transformed-}N(\mu_2, \sigma_2^2, \gamma_2, \omega_2),$$

$$Z_i \mid \theta \sim \text{Bernoulli}(\lambda),$$

where $\theta = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda, \gamma_1, \gamma_2, \omega_1, \omega_2\}$ and the expression "Transformed-N" with four arguments refers to the transformed-normal distribution with the four arguments being the location, scale, power parameter, and "geometric-mean" parameter of the transformed-normal distribution. The complete-data likelihood can be expressed as

$$L(\theta \mid X_1, \ldots, X_n; Z_1, \ldots, Z_n) = \prod_{i=1}^n \left[\frac{1}{\sigma_1\sqrt{2\pi}} e^{-1/2[(\psi(X_i;\gamma_1,\omega_1)-\mu_1)^2/\sigma_1^2]} \left|\left(\frac{X_i}{\omega_1}\right)^{\gamma_1 - 1}\right|\right]^{Z_i} \times \left[\frac{1}{\sigma_2\sqrt{2\pi}} e^{-1/2[(\psi(X_i;\gamma_2,\omega_2)-\mu_2)^2/\sigma_2^2]} \left|\left(\frac{X_i}{\omega_2}\right)^{\gamma_2 - 1}\right|\right]^{(1-Z_i)} J_1 J_2,$$

where J_1 and J_2 are the Jacobians of the scaled transformations $X \to \psi$. If ω_1 and ω_2 were not fixed a priori but instead were the geometric means of the X_i for the respective components, then $J_1 = J_2 = 1$. In our situation, however, because the Z_i's are missing, J_1 and J_2 are functions of $\{X_i\}$, $\{Z_i\}$, and θ, and are not generally equal to 1. Still, J_1 and J_2 are close to 1 when the estimated geometric mean of the sample X_i in component k is close to ω_k. We choose to ignore this minor issue; that is, although we model ω_1 and ω_2 as known from prior considerations, we still assume $J_1 = J_2 = 1$. To do otherwise would greatly complicate our estimation procedure with, we expect, no real benefit; we do not blindly believe such fine details of our model in any case, and we would not expect our procedures to be improved by the extra analytic work and computational complexity.

To keep the distinction clear between the parameters assumed fixed in EM and the parameters being estimated in EM, we partition the parameter into $\theta = \{\theta_{\text{fix}}, \theta_{\text{est}}\}$, where $\theta_{\text{fix}} = \{\gamma_1, \gamma_2, \omega_1, \omega_2, V\}$ and $\theta_{\text{est}} = \{\mu_1, \mu_2, \sigma_1^2, \lambda\}$, and where the variance ratio $V = \sigma_1^2/\sigma_2^2$. Based on this formulation, MLE's of θ_{est} can be obtained from the following EM algorithm:

E step. For $i = 1, \ldots, n$, calculate $Z_i^{(t+1)}$ as in (2), where

$$f_g(X_i \mid \theta^{(t)}) = \frac{1}{\sigma_g\sqrt{2\pi}} e^{-1/2[(\psi(X_i;\gamma_g,\omega_g)-\mu_g)^2/\sigma_g^2]} \left|\left(\frac{X_i}{\omega_g}\right)^{\gamma_g - 1}\right|, \qquad g = 1, 2. \tag{7}$$

M step. Calculate $\mu_1^{(t+1)}$ and $\mu_2^{(t+1)}$ as in (3), $\lambda^{(t+1)}$ as in (5), and $\sigma_1^{2(t+1)}$ and $\sigma_2^{2(t+1)}$ as in (6), with X_i replaced by $\psi(X_i; \gamma_g, \omega_g)$ for $g = 1, 2$; if the variance ratio V were not fixed but were to be estimated, then (4) would be used in place of (6).

2.3.3 Transformed-Normal Mixture Model for Record-Linkage Weights. Let the weights associated with record pairs in a current data set be denoted by W_i, $i = 1, \ldots, n$,

where as before $Z_i = 1$ implies membership in the false-match component and $Z_i = 0$ implies membership in the true-match component. We assume that we have already obtained, from a training sample, (a) values of the power transformation parameters, denoted by γ_F for the false-match component and by γ_T for the true-match component, (b) values of the "geometric mean" parameters in the transformations, denoted by ω_F for the false-match component and by ω_T for the true-match component, and (c) a value for the ratio of the variances between the false-match and true-match components, denoted by V. Our model then becomes

$$W_i|\theta, \{Z_i = 1\} \sim \text{Transformed-}N(\mu_F, \sigma_F^2, \gamma_F, \omega_F),$$

$$W_i|\theta, \{Z_i = 0\} \sim \text{Transformed-}N(\mu_T, \sigma_T^2, \gamma_T, \omega_T),$$

$$Z_i|\theta \sim \text{Bernoulli}(\lambda),$$

where $\theta = \{\mu_F, \mu_T, \sigma_F^2, \sigma_T^2, \lambda, \gamma_F, \gamma_T, \omega_F, \omega_T\}$. We work with $\theta_{fix} = \{\gamma_F, \gamma_T, \omega_F, \omega_T, V = \sigma_F^2/\sigma_T^2\}$ and $\theta_{est} = \{\mu_F, \mu_T, \sigma_F^2, \lambda\}$. The algorithm of Section 2.3.2, with "F" and "T" substituted for "1" and "2," describes the EM algorithm for obtaining MLE's of θ_{est} from $\{W_i; i = 1, \ldots, n\}$ with $\{Z_i; i = 1, \ldots, n\}$ missing and global parameters γ_F, γ_T, ω_F, ω_T, and V fixed at specified values.

2.4 False-Match Rate Estimates and Standard Errors with Fixed Global Parameters

2.4.1 Estimates of the False-Match Rate. Under the transformed-normal mixture model formulation, the false-match rate associated with a cutoff C can be expressed as a function of the parameters θ as

$$\varepsilon(C|\theta) = \frac{\lambda\left[1 - \Phi\left(\frac{\psi_F(C; \gamma_F, \omega_F) - \mu_F}{\sigma_F}\right)\right]}{\lambda\left[1 - \Phi\left(\frac{\psi_F(C; \gamma_F, \omega_F) - \mu_F}{\sigma_F}\right)\right] + (1-\lambda)\left[1 - \Phi\left(\frac{\psi_T(C; \gamma_T, \omega_T) - \mu_T}{\sigma_T}\right)\right]}. \tag{8}$$

Substitution of MLE's for the parameters in this expression provides a predicted false-match rate associated with cutoff C.

Because there is a maximum possible weight associated with perfect agreement in most record-linkage procedures, one could view the weight distribution as truncated above. According to this view, the contribution of the tail above the upper truncation point (say, B), should be discarded by substituting $\Phi([\psi_g(B; \gamma_g, \omega_g) - \mu_g]/\sigma_g)$ for the 1s inside the bracketed expressions (g = F, T as appropriate). Empirical investigation suggests that truncation of the extreme upper tail makes very little difference in predictions. The results in Sections 3 and 4 reflect false-match rate predictions without truncation of the extreme upper tail.

2.4.2 Obtaining an Asymptotic Covariance Matrix for Mixture-Model Parameters From SEM Algorithm. The SEM algorithm (Meng and Rubin 1991) provides a method for obtaining standard errors of parameters in models that are fit using the EM algorithm. The technique uses estimates of the fraction of missing information derived from successive EM iterates to inflate the complete-data variance–covariance matrix to provide an appropriate observed-data variance-covariance matrix. Details on the implementation of the SEM algorithm in our mixture-model setting are deferred to the Appendix.

Standard arguments lead to large-sample standard errors for functions of parameters. For example, the false-match rate $\varepsilon(C|\theta)$ can be expressed as a function of the four components of $\theta_{est} = (\mu_F, \mu_T, \sigma_F^2, \lambda)$ by substituting $\sigma_F/\sqrt{V}$ for σ_T in (8). Then the squared standard error of the estimated false-match rate is given by $SE^2(\varepsilon) \approx \mathbf{d}^T\mathbf{A}\mathbf{d}$, where $\mathbf{A}$ is the covariance matrix for θ_{est} obtained by SEM and the νth component of $\mathbf{d}$ is $d_\nu = \partial\varepsilon/\partial\theta_F$.

2.4.3 Estimates of the Probability of False Match for a Record Pair With a Given Weight. The transformed-normal mixture model also provides a framework for estimating the probability of false match associated with various cutoff weights. To be clear, we draw a distinction between the "probability of false match" and what we refer to as the "neighborhood false-match rate" to avoid any confusion caused by (1) our using a continuous mixture distribution to approximate the discrete distribution of weights associated with a finite number of record pairs, and (2) the fact that there are only finitely many possible weights associated with many record-linkage weighting schemes. The "neighborhood false-match rate around W" is the number of false matches divided by the number of declared matches among pairs of records with composite weights in a small neighborhood of W; with a specific model, the neighborhood false-match rate is the "probability of false match" implied by the relative density of the true-match and false-match components at W.

In terms of the mixture-model parameters, the false-match rate among record pairs with weights between W and $W + h$ is given by

$$\xi(W, h|\theta) = \frac{\lambda p_F(W, h|\theta)}{\lambda p_F(W, h|\theta) + (1-\lambda)p_T(W, h|\theta)},$$

where

$$p_g(W, h|\theta) = \Phi\left(\frac{\psi_g(W + h; \gamma_g, \omega_g) - \mu_g}{\sigma_g}\right) - \Phi\left(\frac{\psi_g(W; \gamma_g, \omega_g) - \mu_g}{\sigma_g}\right), \qquad g = \text{F, T},$$

and $\theta = \{\mu_F, \mu_T, \sigma_F^2, \sigma_T^2, \gamma_F, \gamma_T, \omega_F, \omega_T, \lambda\}$. Although the number of false matches is not a smooth function of the number of declared matches, $\xi(W, h|\theta)$ is a smooth function

of h. The probability of false match under the transformed-normal mixture model is the limit as $h \to 0$ of $\xi(W, h|\theta)$, which we denote as $\eta(W|\theta)$; we obtain

$$\eta(W|\theta) = \lim_{h \to 0} \left\{ \frac{\lambda\left(\frac{\partial p_F(W, h|\theta)}{\partial h}\right)}{\lambda\left(\frac{\partial p_F(W, h|\theta)}{\partial h}\right) + (1-\lambda)\left(\frac{\partial p_T(W, h|\theta)}{\partial h}\right)} \right\} = \frac{\lambda\phi_F^*(W|\theta)}{\lambda\phi_F^*(W|\theta) + (1-\lambda)\phi_T^*(W|\theta)}, \quad (9)$$

where

$$\phi_g^*(W|\theta) = \frac{\lambda}{\sigma_g\sqrt{2\pi}} e^{-1/2(\psi_g(W;\gamma_g,\omega_g)-\mu_g/\sigma_g)^2}\left(\frac{W}{\omega_g}\right)^{\gamma_g - 1}, \quad g = \text{F, T.}$$

Estimates of neighborhood false-match rates are thus routinely obtained by substituting the fixed global parameter values and MLE's of μ_F, μ_T, σ_F, σ_T, and λ into (9).

Because the neighborhood false-match rate captures the trade-off between the number of false matches and the number of declared matches, the problem of setting cutoffs can be cast in terms of the question "Approximately how many declared matches are needed to make up for the cost of a false match?" If subject matter experts who are using a record-linkage procedure can arrive at an answer to this question, then a procedure for setting cutoffs could be determined by selecting a cutoff weight where the estimated neighborhood false-match rate equals the appropriate ratio. Alternatively, one could monitor changes in the neighborhood false-match rate (instead of specifying a "tolerable" neighborhood false-match rate in advance) and could set a cutoff weight at a point just before the neighborhood false-match rate accelerates.

2.5 Reflecting Uncertainty in Global Parameters

When there is more than one source of training data, the information available about both within-site and between-site variability in global parameters can be incorporated into the prior specification. For example, with two training sites, we could combine the average within-site variability in a global parameter with a 1 df estimate of between-site variability to represent prior uncertainty in the parameter. With many sites with covariate descriptors, we could model the multivariate regression of global parameters on covariates.

The procedure we used in the application to census/PES data offers an illustration in the simple case with two training sites available to calibrate a third site. For each of the training-data sites and each of the components (true-match and false-match), joint MLE's were found for $(\gamma_g, \mu_g, \sigma_g^2)$, g = F, T, using a simple grid-search over the power parameters. This yielded two estimates of the power parameters, γ_F and γ_T, and two estimates of the variance ratio V between the false-match and true-match components. Additionally, an estimated variance–covariance matrix for these three parameters was obtained by calculating second differences of the log-likelihood at grid points near the maximum.

Values of each parameter for the mixture-model fitting were drawn from separate truncated-normal distributions with mean equal to the average of the estimates from the two training sites and variance equal to the sum of the squared differences between the individual site parameter values and their mean (i.e., the estimated "between" variance), plus the average squared standard error from the two prior fittings (i.e., the average "within" variance). The truncation ensured that the power parameter for the false-match component was less than 1, that the power parameter for the true-match component was greater than 1, and that the variance ratio was also greater than 1. These constraints on the power parameters were based on the view that because there is a maximum possible weight corresponding to complete agreement and a minimum possible weight corresponding to complete disagreement, the true-match component will have a longer left tail than right tail and the false-match component will have a longer right tail than left tail. The truncation for the variance ratio was based on an assumption that false-match weights will exhibit more variability than true-match weights for these data on the transformed scale as well as on the original scale.

For simplicity, the geometric-mean terms in the transformations (ω_F and ω_T) were simply fixed at the geometric mean of the component geometric means from the two previous sites. If the methods had not worked as well as they did with test and simulated data, then we would have also reflected uncertainty in these parameters.

Due to the structure of our problem, in which the role of the prior distribution is to represent observable variability in global parameters from training data, we presume that the functional form of the prior is not too important as long as variability in global parameters is represented accurately. That is, we anticipate that any one of a number of methods that reflect uncertainty in the parameters estimated from training data will yield interval estimates with approximately the correct coverage properties, i.e., nominal $(1-\alpha) \times 100\%$ interval estimates will cover the true value of the estimand approximately $(1-\alpha) \times 100\%$ or more of the time. Alternative specifications for the prior distribution were described by Belin (1991).

When we fit multiple mixture models to average over uncertainty in the parameters estimated by prior data (i.e., when we use the "full strategy" of Section 2.1), the multiple-imputation framework of Rubin (1987) can be invoked to combine estimates and standard errors from the separate models to provide one inference. Suppose that we fit m mixture models corresponding to m separate draws of the global parameters from their priors and thereby obtain false-match rate estimates $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$ and variance estimates $u_1, u_2, \ldots, u_m$, where $u_i = \text{SE}^2(\varepsilon_i)$ is obtained by the method of Section 2.4.2. Following Rubin (1987, p. 76), we can estimate the false-match rate by

$$\bar{\varepsilon} = \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i$$

and its squared standard error by

$$\mathrm{SE}^2(\bar{\varepsilon}) = \frac{m+1}{m}\left[\frac{1}{m-1}\sum_{i=1}^{m}(\varepsilon_i - \bar{\varepsilon})^2\right] + \frac{1}{m}\sum_{i=1}^{m} u_i .$$

Monte Carlo evaluations documented by Rubin (1987, secs. 4.6–4.8) illustrate that even a few imputations ($m = 2$, 3, or 5) are enough to produce very reasonable coverage properties for interval estimates in many cases. The combination of estimation procedures that condition on global parameters with a multiple-imputation procedure to obtain inferences that average over those global parameters is a powerful technique that can be applied quite generally.

3. PERFORMANCE OF CALIBRATION PROCEDURE ON CENSUS COMPUTER-MATCHING DATA

3.1 Results From Test-Census Data

We use the test-census data described in Section 1.3 to illustrate the performance of the proposed calibration procedure, where determinations by clerks are the best measures available for judging true-match and false-match status. With three separate sites available, we were able to apply our strategy three times, with two sites serving as training data and the mixture-model procedure applied to the data from the third site.

We display results from each of the three tests in Figure 2. The dotted curve represents predicted false-match rates obtained from the mixture-model procedure, with accompanying 95% intervals represented by the dashed curves. Also plotted are the observed false-match rates, denoted by the "O" symbol, associated with each of several possible choices of cutoff values between matches and nonmatches.

We call attention to several features of these plots. First, it is clearly possible to match large proportions of the files with little or no error. Second, the quality of candidate matches becomes dramatically worse at some point where the false-match rate accelerates. Finally, the calibration procedure performs very well in all three tests from the standpoint of providing predictions that are close to the true values and interval estimates that include the true values.

In Figure 3 we take a magnifying glass to the previous displays to highlight the behavior of the calibration procedure in the region of interest where the false-match rate accelerates. That the predicted false-match rate curves bend upward close to the points where the observed false-match rate curves rise steeply is a particularly encouraging feature of the calibration method.

For comparison with the logistic-regression approach, we report in Table 2 (p. 704) the estimated false-match rates across the various sites for records with weights in the interval $[-5, 0]$, which in practice contains both true matches and false matches. Two alternative logistic regression models—one in which $\mathrm{logit}(\eta)$ is modeled as a linear function of matching weight and the other in which $\mathrm{logit}(\eta)$ is modeled as a quadratic function of matching weight, where η is the probability of false match—were fitted to data from two sites to predict false-match rates in the third site. A predictive standard error to reflect binomial sampling variability, as well as uncertainty in parameter estimation, was calculated using

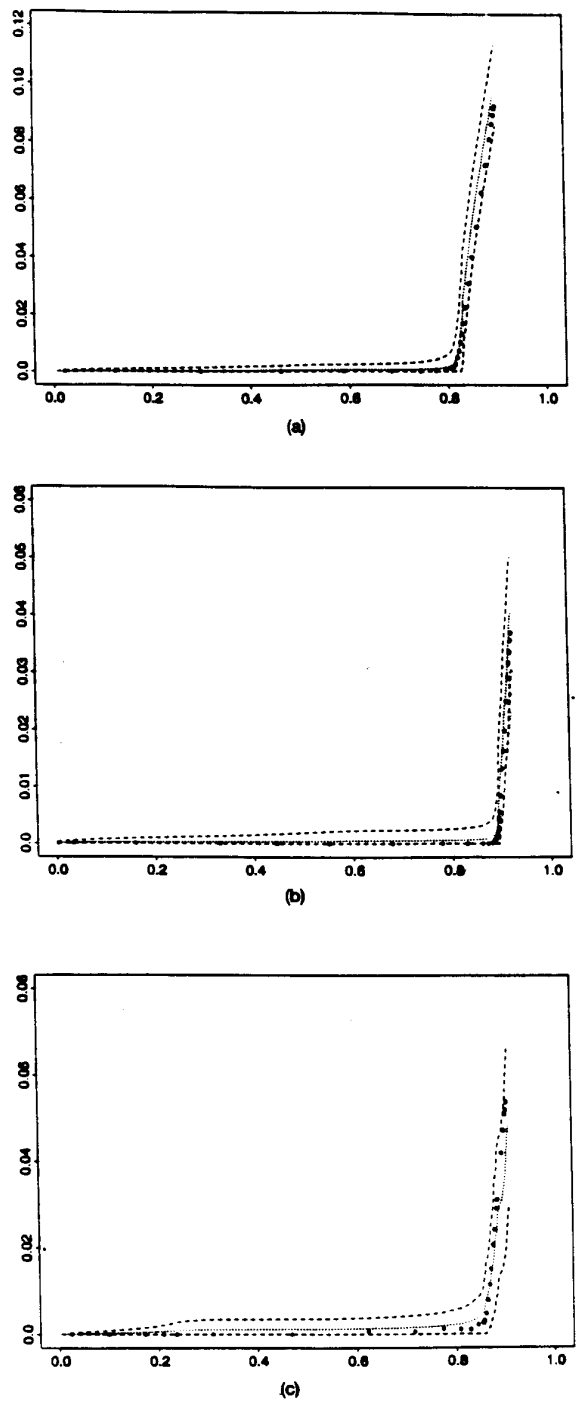

Figure 2. Performance of Calibration Procedure on Test-Census Data: (a) St. Louis, Using Columbia and Washington as Training Sample; (b) Columbia, Using St. Louis and Washington as Training Sample; (c) Washington, Using St. Louis and Columbia as Training Sample. O = observed false-match rate; · · · = predicted false-match rate; --- = upper and lower 95% bounds.

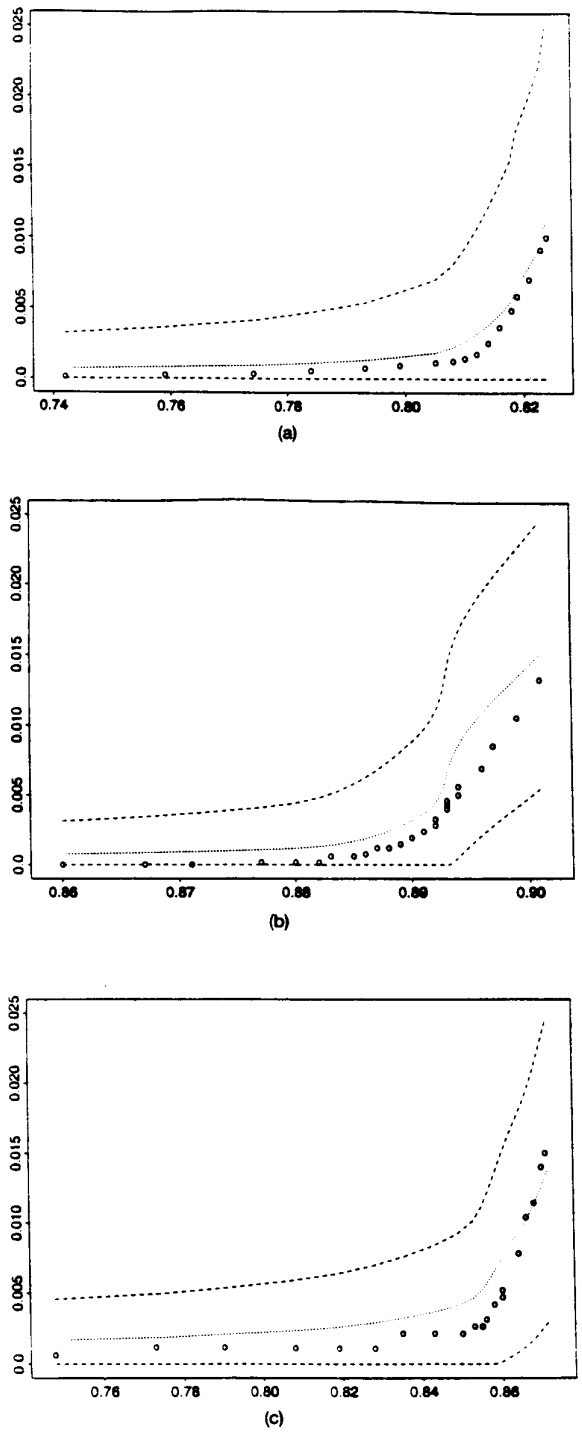

Figure 3. Performance of Calibration Procedure in Region of Interest: (a) St. Louis, Using Columbia and Washington as Training Sample; (b) Columbia, Using St. Louis and Washington as Training Sample; (c) Washington, Using St. Louis and Columbia as Training Sample. O = observed false-match rate; · · · = predicted false-match rate; --- = upper and lower 95% bounds.

$$\mathrm{SE} = \frac{1}{n^*}\sqrt{\sum_{i=1}^{n^*} \hat{\eta}_i(1-\hat{\eta}_i)[1+\mathbf{w}_i^T \mathrm{var}(\hat{\beta})\mathbf{w}_i]},$$

where n^* is the number of record pairs with weights in the target interval and $\hat{\eta}_i$ is the predicted probability of false match for candidate pair i. For the linear logistic model, we have $\mathbf{w}_i = (1, w_i)^T$, where $\mathbf{w}_i$ is the weight for record pair i and $\mathrm{var}(\hat{\beta})$ is the 2×2 covariance matrix of the regression parameters, whereas for the quadratic logistic model we have $\mathbf{w}_i = (1, w_i, w_i^2)^T$, where $\mathrm{var}(\hat{\beta})$ is the 3×3 covariance matrix of the regression parameters.

As can be seen from Table 2, the predicted false-match rates from two alternative logistic regression models often were not in agreement with the observed false-match rates; in fact, they were often several standard errors apart. Because weights typically depend on site-specific data, this finding was not especially surprising. It is also noteworthy that the estimate of the quadratic term β_2 in the quadratic models was more than two standard errors from zero using the St. Louis data ($p = .029$) but was near zero using the Columbia and Washington data sets individually ($p = .928$ and $p = .719$). Using the mixture-model calibration approach, in two of the three sites the observed false-match rate is within two standard errors of the predicted false-match rate, and in the other site (St. Louis) the mixture-model approach is conservative in that it overstates the false-match rate. We regard this performance as superior to that of logistic regression—not surprising in light of our earlier discussion in Section 1.3 of why we eschewed logistic regression in this setting. Refining the mixture-model calibration approach through a more sophisticated prior distribution for global parameters (e.g., altering the prior specification so that urban sites are exchangeable with one another but not with rural sites) may result in even better performance by reflecting key distinctions in distributional shapes across sites.

3.2 A Limitation in the Extreme Tails

For small proportions of the records declared matched, counter-intuitive false-match rate estimates arise, with false-match rate estimates increasing as the proportion declared matched goes to zero. Such effects are explained by the false-match component being more variable than the true-match component, so that in the extreme upper tail of the component distributions the false-match component density is greater than the true-match component density. To avoid nonsensical results (since we know that the false-match rate should go to zero as the weight approaches the maximum possible weight), we find the points along the false-match-rate curve and the upper interval-estimate curve, if any, where the curves depart from a monotone decreasing pattern as the proportion declared matched approaches zero. From the point at which the monotone pattern stops, we linearly interpolate false-match rate estimates between that point and zero. We are not alarmed to find that the model does not fit well in the far reaches of the upper tails of component distributions, and other smoothing procedures may be preferred to the linear-interpolation procedure used here.

Table 2. Performance of Mixture-Model Calibration Procedure on Test-census Matching Weights in the Interval [−5, 0]

Site to be predicted	Observed false-match rate among cases with weights in [−5, 0]	Predicted false-match rate (SE) for cases with weights in [−5, 0] under linear logistic model; that is logit(η) = $\alpha + \beta$ (Wt)	Predicted false-match rate (SE) for cases with weights in [−5, 0] under quadratic logistic model; that is logit(η) = $\alpha + \beta_1$ (Wt) + β_2 (Wt)^2	Predicted false-match rate (SE) for cases with weights in [−5, 0] based on mixture-model calibration method
St. Louis	.646 (= 73/113)	.417 (.045)	.429 (.045)	.852 (.033)
Columbia	.389 (= 14/36)	.584 (.079)	.613 (.079)	.524 (.083)
Washington	.294 (= 5/17)	.573 (.115)	.597 (.115)	.145 (.085)

4. SIMULATION EXPERIMENT TO INVESTIGATE PROPERTIES OF CALIBRATION PROCEDURE

4.1 Description of Simulation Study

Encouraged by the success of our method with real data, we conducted a large simulation experiment to enhance our understanding of the calibration method and to study statistical properties of the procedure (e.g., bias in estimates of the false-match rate, bias in estimates of the probability of false match, coverage of nominal 95% interval estimates for false-match rates, etc.) when the data generating mechanism was known. The simulation procedure involved generating data from two overlapping component distributions, with potential "site-to-site" variability from one weight distribution to another, to represent plausible weights, and replicating the calibration procedure on these samples.

Beta distributions were used to represent the components of the distribution of weights in the simulation experiment. Simulated weights thus were generated from component distributions that generally would be skewed and that have a functional form other than the transformed-normal distribution used in our procedure. The choice of beta-distributed components was convenient in that simple computational routines were available (Press, Flannery, Teukolsky, and Vetterling 1986) to generate beta-random deviates and to evaluate tail probabilities of beta densities.

Table 3 lists factors and outcome variables that were included in the experiment. Here we report only broad descriptive summaries from the simulation study. Greater detail on the design of the experiment and a discussion of the strategy for relating simulation outcomes to experimental factors have been provided by Belin (1991). The calibration procedure was replicated 6,000 times, with factors selected in a way that favored treatments that were less costly in terms of computer time (for details, again see Belin 1991).

4.2 Results

Figure 4 displays the average relative bias from the various simulation replicates in a histogram. Due to the way that we have defined relative bias (see Table 3), negative values correspond to situations where the predicted false-match rate is greater than the observed false-match rate; that is, negative relative bias corresponds to conservative false-match rate estimates. It is immediately apparent that the calibration procedure is on target in the vast majority of cases, with departures falling mostly on the "conservative" side, where the procedure overstates false-match rates and only a few cases where the procedure understates false-match rates.

The few cases in which the average relative bias was unusually large and negative were examined more closely, and all of these had one or more cutoffs where the observed false-match rate was zero and the expected false-match rate was small. In such instances the absolute errors are small, but relative errors can be very large. Clearly, however, errors between a predicted false-match rate of .001 or .002 and observed false-match rate of 0 presumably are not of great concern in applications.

There was a single replicate that had a positive average relative bias substantially larger than that of the other replicates. Further investigation found that a rare event occurred in that batch, with one of the eight highest-weighted records being a false match, which produced a very high relative error. In this replicate, where the predicted false-match rate under the mixture model was .001, the observed false-match rate was .125 and the expected (beta) false-match rate was .00357. Belin (1991) reported that other percentiles of the distribution in this replicate were fairly well calibrated (e.g., predicted false-match rates of .005, .01, .10, .50, and .90 corresponded to expected beta false-match rates of .007, .011, .130, .455, and .896); thus it was apparently a rare event rather than a breakdown in the calibration method that led to the unusual observation.

With respect to the coverage of interval estimates, we focus on simulation results when the SEM algorithm was applied to calculate a distinct covariance matrix for each fitted mixture model (n = 518). In the other simulation replicates, shortcuts in standard error calculations were taken so as to avoid computation; Belin (1991) reported that these shortcut methods performed moderately worse in terms of coverage. For nominal two-sided 95% intervals, the calculated intervals covered observed false-match rates 88.58% of the time (SE = 1.94%); for nominal one-sided 97.5% intervals, the calculated intervals covered observed false-match rates 97.27% of the time (SE = 1.45%). Thus the calibration method does not achieve 95% coverage for two-sided interval estimates, but when it errs it tends to err on the side of *overstating* false-match rates, so that the one-sided interval estimates perform well.

Table 3. Description of Factors and Outcomes in Simulation Experiment

Experimental factors	*Comments*
1. Number of sources of training data	1. Possible values = {3, 4, . . . , 30}
2. Sizes of training samples	2–3. Possible values = {2,000, 2,001, . . . , 9,999}
3. Size of current data base	4. Possible values in [.01, .5]
4. Mixing proportion between false-match and true-match components	5. $Be(a_F, b_F)$, $a_F \sim U(1.75, 2)$ $b_F \sim U(3, 6)$
5. Shape of the false-match component	6. $Be(a_T, b_T)$, $a_T \sim U(8, 12)$ $b_T \sim U(1.75, 2)$
6. Shape of true-match component	7–8. Separate draws from 5–6 for separate sites, or common shapes across sites
7. Amount of site-to-site variability in mixing proportion	9. Perform SEM for each mixture model being fit, or a short-cut method to save computing time based on approximations
8. Amount of site-to-site variability in shapes of component distributions	10. Possible values = {3, 5, 10}
9. Method for calculating standard errrors	
10. Number of mixture models fit	
Outcomes measured in simulation	*Comments*
1. Average relative bias in false-match rate estimates across 20 prespecified predicted false-match rates	1. Prespecified false-match rates = {.001, .002, . . . , .015, .02, .025, .03, .04, .05}
2. Average two-sided coverage rate = average of 20 indicators for whether interval estimates of false-match rate covered observed false-match rate	Relative bias = [(observed false-match rate) − (predicted false-match rate)]/√(expected false-match rate)
3. Average one-sided coverage rate = average of 20 indicators for whether interval estimates of false-match rate covered or overstated false-match rate	"predicted false-match rate" = estimated false-match rate calculated under transformed-normal mixture model
4. Expected probability of false match for 10 predicted probabilities of false match	"observed false-match rate" = {# false matches}/{# declared matches} at given cutoff
	"expected false-match rate" = {tail area of Beta false-match component}/{sum of tail areas of Beta component distributions}
	2–3. Same prespecified false-match rates as in 1.
	4. Estimated probabilities of false match = {.00125, .0025, .005, .01, .02, .10, .25, .50, .75, .90}

Turning to the performance of the estimated probabilities of false match (i.e., neighborhood false-match rates) obtained from the fitted mixture models, Table 4 provides the mean, standard deviation, minimum, and maximum of the true underlying probabilities being estimated by the calibration procedure. Although in specific sites the calibration procedure substantially understates or overstates false-match rates, the procedure appears to have good properties in the aggregate.

5. DISCUSSION

Previous attempts at estimating false-match rates in record linkage were either unreliable or too cumbersome for practical use. Although our method involves fitting nonstandard models, other researchers have used software that we developed to implement the technique in at least two other settings (Ishwaran, Berry, Duan, and Kanouse 1991; Scheuren and Winkler 1991, 1993). This software is available on request from the first author.

Analyses by Belin (1991) have revealed that the deficiencies in the calibration procedure typically occurred where the split in the proportion of records between the two components was very extreme. For example, after excluding a few dozen replicates where 99% or more of the records were declared matched above the point where the procedure predicted a false-match rate of .005, there was no evidence that sample sizes of the data bases being matched had an impact on the accuracy of estimated probabilities of false match, implying that breakdown of the calibration procedure appears to be a threshold phenomenon.

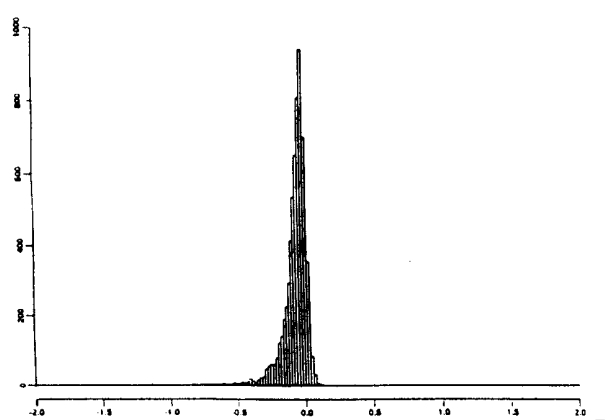

Figure 4. Histogram of Average Relative Bias Across Simulation Replicates.

Table 4. Performance of Estimated Probabilities of False-Match in Predicting True Underlying Probabilities

Estimated probability	*Mean of actual probabilities*	*Std. deviation of actual probabilities*	*Minimum of actual probabilities*	*Maximum of actual probabilities*
.00125	.00045	.00075	.00000	.0105
.0025	.0012	.00143	.00000	.0170
.005	.0030	.00266	.00000	.0295
.01	.0072	.00470	.00002	.059
.02	.0160	.00818	.00019	.1183
.10	.0813	.03231	.00155	.4980
.25	.2086	.07385	.00549	.8094
.50	.4555	.11996	.02812	.9551
.75	.7459	.10768	.20013	.9988
.90	.9244	.05222	.56845	1.000

Finally, on several occasions when we have discussed these techniques and associated supporting evidence, we have been questioned about the validity of using determinations of clerks as a proxy for true-match or false-match status. Beyond pointing to the results of the simulations, we note that (a) clerical review is as close to truth as one is likely to get in many applied contexts, and (b) possible inaccuracy in assignment of match status by clerks is no criticism of the calibration procedure. This methodology provides a way to calibrate false-match rates to whatever approach is used to identify truth and falsehood in the training sample and appears to be a novel technique that is useful in applied contexts.

APPENDIX: IMPLEMENTATION OF THE SEM ALGORITHM

The SEM algorithm is founded on the identity that the observed-data observed information matrix $\mathbf{I}_{obs}$ for a k-dimensional parameter θ can be expressed in terms of the conditional expectation of the "complete-data" observed information matrix evaluated at the MLE ($\mathbf{I}_{com}$) as

$$\mathbf{I}_{obs} = (\mathbf{I} - \mathbf{DM})\mathbf{I}_{com},$$

where $\mathbf{I}$ is the $k \times k$ identity matrix and $\mathbf{DM}$ is the Jacobian of the mapping defined by EM (i.e., of the mapping that updates parameter values on the tth iteration to those on the $(t+1)$st iteration) evaluated at the MLE $\hat{\theta}$. Taking inverses of both sides of the previous equation yields the identity

$$\mathbf{V}_{obs} = \mathbf{I}_{obs}^{-1} = \mathbf{V}_{com} + \Delta\mathbf{V},$$

where $\mathbf{V}_{com} = \mathbf{I}_{com}^{-1}$ and $\Delta\mathbf{V} = \mathbf{V}_{com}\mathbf{DM}(\mathbf{I} - \mathbf{DM})^{-1}$, the latter reflecting the increase in variance due to the missing information.

The SEM procedure attempts to evaluate the $\mathbf{DM}$ matrix of partial derivatives numerically. First, EM is run to convergence and the MLE $\hat{\theta}$ is obtained. To compute partial derivatives, all but one of the components of the parameter are fixed at their MLE's, and the remaining component is set to its value at the tth iteration, say $\theta^{(t)}(i)$. Then, after taking a "forced EM step" by using this parameter value as a start to a single iteration (E step and M step), the new estimates, say $\tilde{\theta}_j^{(t+1)}(i)$ for $j = 1, \ldots, k$, yield the following estimates of partial derivatives:

$$r_{ij}^{(t)} = \frac{\tilde{\theta}_j^{(t+1)}(i) - \hat{\theta}_j}{\theta_i^{(t)}(i) - \hat{\theta}_i}.$$

It is necessary to perform k forced EM steps at every iteration of SEM—although Meng and Rubin (1991) pointed out that once convergence is reached for each component of the vector $(r_{i'1}, r_{i'2}, \ldots, r_{i'k})$, it is no longer necessary to perform the forced EM step for component $i = i'$.

Because we regard the variance ratio as fixed when fitting our mixture models, we are actually estimating four parameters in the calibration mixture-model setting: the locations of the two components, one unknown scale parameter, and a mixing parameter. We can calculate the complete-data information matrix for $(\mu_F, \mu_T, \sigma_F^2, \lambda)$ as

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & a_{13} & 0 \\ 0 & a_{22} & a_{23} & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ 0 & 0 & 0 & a_{44} \end{pmatrix},$$

where

$$a_{11} = \frac{\sum_{i=1}^n Z_i}{\sigma_F^2} \qquad a_{22} = \frac{V\sum_{i=1}^n (1 - Z_i)}{\sigma_F^2} \qquad a_{33} = \frac{\sum_{i=1}^n Z_i}{\sigma_F^6}$$

$$a_{44} = \frac{\sum_{i=1}^n Z_i}{\lambda^2} + \frac{\sum_{i=1}^n (1 - Z_i)}{(1-\lambda)^2}$$

$$a_{31} = a_{13} = \frac{\sum_{i=1}^n Z_i(w_i - \mu_F)}{\sigma_F^4}$$

$$a_{32} = a_{23} = \frac{V\sum_{i=1}^n (1 - Z_i)(w_i - \mu_T)}{\sigma_F^4}.$$

The missing information in our problem arises from the fact that the Z_i's are unknown.

Because the covariance matrix is 4×4, every iteration of the SEM algorithm takes roughly four times as long as an iteration of EM. It also should be pointed out that the SEM algorithm relies on precise calculation of MLE's. Although it may only be necessary to run EM for 10 or 20 iterations to obtain accuracy to two decimal places in MLE's, it might take 100 or more iterations to obtain accuracy to, say, six decimal places. These aspects of the SEM algorithm can make it computationally expensive.

The $\mathbf{DM}$ matrix containing the r_{ij}'s will not generally be symmetric, but of course the resulting covariance matrix should be symmetric. If the resulting covariance matrix is not symmetric even though several digits of numerical precision are obtained for the MLE and the r_{ij}'s, this reflects an error in the computer code used to implement the forced SEM steps or perhaps in the code for the E step and M step themselves. The symmetry or lack thereof in the resulting covariance matrix thus provides a diagnostic check on the correctness of the program.

Experience with the SEM algorithm suggests that convergence of the numerical approximations to the partial derivatives of the mapping often occurs in the first few iterations and further reveals that beyond a certain number of iterations, the approach can give nonsensical results owing to limitations in numerical precision, just as with any numerical differentiation procedure. Meng and Rubin (1991) suggested specifying a convergence criterion for the r_{ij}'s and ceasing to calculate these terms once the criterion is satisfied for all $j = 1, \ldots, k$. An alternative (used in producing the results in Secs. 3 and 4) involves running the SEM algorithm for eight iterations, estimating all partial derivatives of the mapping on each iteration, assessing which two of the eight partial derivative estimates are closest to one another, and taking the second of the two as our estimate of the derivative. Experience with this approach suggests that it yields acceptable results for practice in that the off-diagonal elements of the resulting covariance matrix agree with one another to a few decimal places.

[Received February 1993. Revised November 1993.]

REFERENCES

Aitkin, M., and Rubin, D. B. (1985), "Estimation and Hypothesis Testing in Finite Mixture Models," *Journal of the Royal Statistical Society*, Ser. B, 47, 67–75.

Belin, T. R. (1989a), "Outline of Procedure for Evaluating Computer Matching in a Factorial Experiment," unpublished memorandum, U.S. Bureau of the Census, Statistical Research Division.

——— (1989b), "Results from Evaluation of Computer Matching," unpublished memorandum, U.S. Bureau of the Census, Statistical Research Division.

——— (1990), "A Proposed Improvement in Computer Matching Techniques," in *Statistics of Income and Related Administrative Record Research: 1988–1989*, Washington, DC: U.S. Internal Revenue Service, pp. 167–172.

——— (1991), "Using Mixture Models to Calibrate Error Rates in Record-Linkage Procedures, with Application to Computer Matching for Census Undercount Estimation," Ph.D. thesis, Harvard University, Dept. of Statistics.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 26, 206–252.

Brown, P., Laplant, W., Lynch, M., Odell, S., Thibaudeau, Y., and Winkler, W. (1988), *Collective Documentation for the 1988 PES Computer-Match Processing and Printing*, Vols. I–III, Washington, DC: U.S. Bureau of the Census, Statistical Research Division.

Copas, J., and Hilton, F. (1990), "Record Linkage: Statistical Models for Matching Computer Records," *Journal of the Royal Statistical Society*, Ser. A, 153, 287–320.

Dawid, A. P. (1976), "Properties of Diagnostic Data Distributions," *Biometrics*, 32, 647–658.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Efron, B. (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of the American Statistical Association*, 70, 892–898.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183–1210.

Halperin, M., Blackwelder, W. C., and Verter, J. I. (1971), "Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches," *Journal of Chronic Diseases*, 24, 125–158.

Hogan, H. (1992), "The 1990 Post-Enumeration Survey: An Overview," *The American Statistician*, 46, 261–269.

Ishwaran, H., Berry, S., Duan, N., and Kanouse, D. (1991), "Replicate Interviews in the Los Angeles Women's Health Risk Study: Searching for the Three-Faced Eve," technical report, RAND Corporation.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414–420.

Kelley, R. P. (1986), "Robustness of the Census Bureau's Record Linkage System," in *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 620–624.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

Maclean, C. J., Morton, N. E., Elston, R. C., and Yee, S. (1976), "Skewness in Commingled Distributions," *Biometrics*, 32, 695–699.

Mantel, N., and Brown, C. (1974), "Alternative Tests for Comparing Normal Distribution Parameters Based on Logistic Regression," *Biometrics*, 30, 485–497.

Meng, X. L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance–Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909.

Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford, U.K.: Oxford University Press.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, 130, 954–959.

Orchard, T., and Woodbury, M. A. (1972), "A Missing Information Principle: Theory and Applications," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 697–715.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986), *Numerical Recipes: The Art of Scientific Computing*, Cambridge, U.K.: Cambridge University Press.

Rogot, E., Sorlie, P. D., and Johnson, N. J. (1986), "Probabilistic Methods in Matching Census Samples to the National Death Index," *Journal of Chronic Diseases*, 39, 719–734.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Scheuren, F., and Winkler, W. E. (1991), "An Error Model for Regression Analysis of Data Files That are Computer Matched," in *Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census*, pp. 669–687.

——— (1993). "Regression Analysis of data files that are computer matched," *Siney Methodology*, 19, 39–58.

Tepping, B. J. (1968), "A Model for Optimum Linkage of Records," *Journal of the American Statistical Association*, 63, 1321–1332.

Thibaudeau, Y. (1989), "Fitting Log-Linear Models in Computer Matching," in *Proceedings of the American Statistical Association, Section on Statistical Computing*, pp. 283–288.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: John Wiley.

Weisberg, S. (1980), *Applied Linear Regression*, New York: John Wiley.

Winkler, W. E. (1989), "Near-Automatic Weight Computation in the Fellegi–Sunter Model of Record Linkage," in *Proceedings of the Fifth Annual Research Conference, U.S. Bureau of the Census*, pp. 145–155.