This is an quite an indictment and should be accompanied by convincing proof of unethical behavior and distortion of facts. Belin and Rolph have not come up with a single fact that I have been careless with, nor any instance of unethical behavior. On the other hand, they have gotten their facts wrong, have been careless in reading, have found contradictions and obfuscations where there are none and have spent most of their time on irrelevant side issues and morality mongering.

Rejoinder

D. Freedman and K. Wachter

1. INTRODUCTION

Census adjustment is not an easy topic. We are grateful to the discussants for their efforts at clarifying the issues. One other idea will not be controversial: Rob Kass and Ram Gnanadesikan deserve thanks for putting this exchange together and bringing it to a successful conclusion.

The commentaries fall naturally into two groups, those from outside the United States and those from inside. It is valuable to have perspectives gained from experience in other countries. We marvel, naturally, at errors measured in hundreds, which Lyberg and Lundstrom attribute to Sweden's PIN-keyed registers. Australia as described by Steel makes an interesting contrast to Britain (Diamond and Skinner), in terms of the trust accorded to results from demographic analysis in Britain and the distrust in Australia—even though Australia has effective monitoring of international migration, which removes one of the chief components of uncertainty in demographic analysis for the United States.

Belin and Rolph (BR) have a free-spirited and wide-ranging commentary which reviews many previous exchanges on the census. Much as we like the authors, we differ with them on readings of the technical and historical record and on matters of scientific principle. With respect to the census, Ericksen, Fienberg and Kadane (EFK) are among the oldest and most familiar of our opponents; but on this occasion, as we shall explain, their critique is off the mark entirely.

According to the rules of engagement, we do not comment on BR's rejoinder and they do not comment on ours, so we get the last word in this exchange—on these pages of this journal. Silence cannot be interpreted as consent: we are sure that BR and EFK will continue the argument in some other forum.

The Census Bureau's latest thinking on the 1991–1992 adjustments is described in Fay and Thompson (1993). Our discussants frequently refer to this paper for arbitration, and we shall too. We hope Bob Fay and John Thompson will not mind such close textual analysis.

Despite the scope of BR's remarks, our paper was not really about the bottom-line question: whether adjustment would have made errors in state population shares better or worse. It was, rather, about a "wild card" in the Census Bureau's assessments of state and local coverage error: heterogeneity. Heterogeneity was omitted from the bureau's loss function analysis. Statisticians on all sides have been arguing ever since what kind of difference that could have made.

In our paper, we measured the difference that heterogeneity does make, in a context that allows an exact answer. We used the same loss function that the Bureau did, with proxy variables instead of undercounts. In our context, the adjustment factors are known with perfect accuracy, so that errors of adjustment are due purely to heterogeneity, and loss itself can be calculated. We found the following:

1. The omission of heterogeneity does bias the estimated risks.
2. Depending on the proxy, the bias can be small or it can be large.
3. The bias can go either way, for or against adjustment.

In particular, we established that loss function analysis can be strongly biased in favour of adjustment; EFK and BR react quite critically to this finding. Before we answer them, let us recall the larger picture in which such arguments take their place.

2. BACKGROUND

Would the proposed adjustment of the 1990 census, or of the intercensal estimates, have improved the accuracy of population shares held by the various states? "Loss function analysis" attempts to balance errors in the census against errors in adjustment, and it seems to be the principal statistical argument that adjustment would improve on the census (BR,

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science*. www.jstor.org

Section 2; Woltman et al., 1991; Mulry and Spencer, 1993; Freedman, Wachter, Cutler and Klein, 1994).

Loss function analysis has several ingredients. The first is estimated sampling error. In the context of census adjustment, the Bureau's estimates for variance (based on their smoothing model) were substantially too optimistic, by a factor of 2 or 3 (Fay and Thompson, 1993, page 83; Freedman et al., 1993); this problem disappears for the intercensals, since the smoothing model was not used. Another input to loss function analysis is the estimated levels of bias in the PES (due to matching error, census day address error and so forth).

The Bureau elected not to estimate bias in the PES for geographical areas, or even for poststrata. Instead, errors were determined for very large aggregates of poststrata, called evaluation strata. With the census, there were 1,392 poststrata and 13 evaluation strata; with the intercensals, there were 357 poststrata and 10 evaluation strata. Breiman shows the Census Bureau's 1991 estimates for bias to be far too optimistic. There seems to be some agreement on these points, the discussants notwithstanding (Section 6 below).

To adjust population counts in small areas, including states, the bureau made the "synthetic assumption": undercount rates are constant within poststrata across geography. Thus, rates are determined by demographics not geography. Failures in the synthetic assumption are termed "heterogeneity." The loss function analysis was done on the basis of the synthetic assumption and therefore could not measure the impact of heterogeneity. To allocate biases from evaluation strata to poststrata, error rates were assumed constant within evaluation strata across poststrata, and then within poststrata across geography—an even stronger form of the synthetic assumption. (For details, see Freedman, Wachter, Cutler and Klein, 1994, pages 260–261.)

Section 5 below indicates how the results of loss function analysis change if we correct the variance and bias estimates. We also show that the scheme for allocating errors from evaluation strata to individual poststrata has considerable influence (Freedman, Wachter, Cutler and Klein, 1994, pages 262, 264). Our present paper, however, dealt with another topic in the adjustment debate: the impact of heterogeneity on loss function analysis. Does heterogeneity create bias in estimated risks? ("Risk" is expected loss, and the goal of the loss function analysis was to obtain unbiased estimates of risk.)

The modelers have taken several positions:

1. Heterogeneity is trivial.
2. If not, the impact on loss function analysis is trivial ("robustness").
3. If neither 1 nor 2 is admitted, then loss function analysis is "conservative": the proadjustment position is even stronger than the data say it is.

Belin and Rolph now concede that heterogeneity is substantial, but stand on point 2; EFK concede nothing, but insist on point 3. The reasoning will be discussed below. We stress that the issue is bias in estimated risks, not bias or variance in estimated counts—a distinction that some discussants seem to have found quite subtle.

Our paper uses proxies for the undercount: variables (such as substitutions or imputations) thought by the Census Bureau to resemble the undercount with respect to heterogeneity. The basic setup involves a matrix; the row index i corresponds to poststrata, running from 1 to 357 in our examples; the column index j corresponds to geographical areas, for instance, the 50 states and Washington, D.C. In each cell, we have the census count c_{ij} and the proxy undercount t_{ij}. The "true" population of area j is the column sum of $c_{ij} + t_{ij}$, with i running from 1 to 357.

Take this from the perspective of statisticians who know the census counts c_{ij} in each cell, but not the t_{ij}. Although the undercounts are unknown, row totals are given:

$$\sum_{j=1}^{51} t_{ij} \quad \text{for } i = 1, \ldots, 357.$$

Now the t_{ij} can be estimated by the "synthetic method,"

$$\hat{t}_{ij} = f_i \times c_{ij}.$$

The factors f_i are computed as follows:

$$f_i = \frac{\sum_{j=1}^{51} t_{ij}}{\sum_{j=1}^{51} c_{ij}}.$$

The row totals of the proxy undercounts are given, so the factors f_i are free of error. (We do not believe that all our discussants paid due attention to this point.) In the adjustment literature, $1 + f_i$ is an "adjustment factor" which adjusts the census count in a cell to its estimated true count.

The "adjusted" population of area j can be obtained as a column total of $c_{ij} + \hat{t}_{ij}$, and shares can be computed in the obvious way. Moreover, the squared error for the raw census and the adjustment can be estimated using the synthetic assumption; comparing these estimates to the true errors enables us to measure the bias in loss function analysis due to heterogeneity.

Risks are estimated as follows. On the basis of the synthetic assumption, adjustment "must" get the

right answer, so its squared error is zero; and the squared error of the census can be calculated with the adjusted population taken as "truth." Of course, in the presence of heterogeneity, such estimates may be quite misleading.

The setup is rather like real census adjustment. The whole object of the PES, poststratification, smoothing and so forth is to obtain estimates of adjustment factors $1 + f_i$. However, our setup differs from adjustment in two major respects: (i) The row totals of the proxies (and hence the adjustment factors themselves) are free of error. (ii) We are working with proxies rather than undercounts. The advantage of (i) is that we can focus on heterogeneity, pure and simple; (ii) is the price.

3. ERICKSEN, FIENBERG AND KADANE

The chief technical argument offered by EFK is that "Heterogeneity... is not inimical to adjustment," and they cite theorems to support their position. However, EFK have utterly missed the point. Their theorems say that, in some circumstances, estimated adjustment factors are too small. In our context, the adjustment factors are not estimated. They are known exactly. There is zero bias and zero variance. The theorems cited by EFK are irrelevant to the calculations that we present. EFK also complain that our adjustment factors are not smoothed. Of course not—smoothing would (if all went well) reduce sampling error in the factors. Our factors are known and are not subject to sampling error. Smoothing them would just add bias.

Even in a context where adjustment factors are only estimates, a distinction must be drawn (EFK always resist this) between (i) adding more people to rows of the matrix and (ii) changing the population shares of columns. The mathematics cited by EFK suggest that, in some circumstances, people should be added to a poststratum beyond what is done by the dual system estimator; in this sense, adjustment is "conservative." That, however, has no direct implications for population shares of states, because the theorems do not say where the additional people live. State shares are the focus of the analysis, not population counts for poststrata.

When discussing our proxy "DIFF," EFK impute to us a regression model for undercount rates, and demand justification. "What evidence do we have for believing such models," they ask. They may have woken up to the idea that assumptions need to be justified—a welcome development. However, they seem to misunderstand DIFF. The Census Bureau's proxies all have a somewhat unrealistic feature: they are positive everywhere. We constructed DIFF, not by regression, but to have a positive part like gross omissions (the undercounts) and a negative part like erroneous enumerations (overcounts). Indeed, the correlations reported in our Table 10 should warn against regression modeling.

Proxies are used (by the Bureau and by us) not to model undercounts in the sense of regression, but as *proxies*: variables that are analogous to undercounts and that stand in for them (Fay and Thompson, 1993, Section 4.3). If the analogy is a bad one (heterogeneity in undercounts does not behave like heterogeneity in the proxies), then the Bureau's research effort on the proxies casts no light on adjustment, and neither does our paper. We made the point earlier; it bears repetition. Conversely, if heterogeneity in undercounts behaves like heterogeneity in the proxies, our work says something about loss function analysis: appreciable bias in risk estimates is a distinct possibility.

Ericksen, Fienberg and Kadane do have one serious point. The 1,392 poststrata used for adjusting the census are different from the 357 for the intercensals. Our findings on the 357 poststrata are most relevant to the bureau's loss function analyses on the intercensals (Bureau of the Census, 1992c, 1993). According to EFK, the fact that "the 357 poststrata have too much residual variability...comes as no surprise"; indeed, dropping the number of poststrata from 1,392 to 357 "was sure to introduce greater heterogeneity." Since we have looked at heterogeneity for the 1,392 poststrata, EFK's theory can be tested.

Results for the 1,392 poststrata are based on a sample covering 200,000 blocks, drawn from the 1990 census by the bureau for its P12 Evaluation Project (Kim, 1991). Unlike calculations for the 357 poststrata, which were based on the whole census and therefore not subject to sample variability, our estimates for the 1,392 poststrata include a correction term for sampling error (Wachter and Freedman, 1994). Results are shown in Table R1; the proxies are substitutions, allocations, multiunit housing and non mailbacks.

The first two columns correspond to columns 3 and 4 in Table 5 of our paper. The last column measures heterogeneity across local areas rather than states. The standard deviations within poststrata across states are all of roughly the same magnitude as before. The value for multiunit housing is bigger, the others are smaller.

Ericksen, Fienberg and Kadane may be right, that 1,392 poststrata do have less residual heterogeneity than 357, although the difference seems minor. Their main idea, that the levels of heterogeneity uncovered by our paper arise from dropping the number of poststrata from 1,392 to 357, turns out to be mistaken. Even with 1,392 poststrata, there was plenty of het-

TABLE R1
Heterogeneity with 1,392 poststrata

	Standard Deviations (%)		
	Across post-strata	Within poststrata across states	Within post-strata across local areas
SUB	0.7	0.6	2.3
ALL	8.0	2.9	7.1
MUH	23.7	10.4	22.3
NMB	12.0	4.3	10.7

Notes: "Local" areas have populations of about 10,000; r.m.s. standard deviations across poststrata are reported, with 12 Indian poststrata excluded. Standard errors are relatively small. For additional details, see Wachter and Freedman (1994).

erogeneity at the state level, while the level of heterogeneity for substate areas was striking.

4. BELIN AND ROLPH

Model Validation

As we read them, BR simply do not accept the idea that modelers have a responsibility to validate the models (Section 2.3). The closest they come is the idea that models are better when "violations of the assumptions ... appear to be harder to find" (Section 2.4). Once again, the burden of disproof is placed on the critic.

Loss Function Analysis

Belin and Rolph ask (Section 2.6), "What is the right loss function?" This a question without an answer, because loss functions do not measure real losses; they are only summaries of error distributions, which may or may not be useful. The papers cited by BR at the end of Section 2.6 do not face up to that central difficulty. For example, here is Zaslavsky (1993a, page 1092), arguing for squared errors: "If there is a social good (such as a government expenditure) to be allocated, and aggregate utility is a smooth (twice differentiable) function...." But why is there an aggregate utility function in the first place, rough or smooth? Over the years, the attempt to construct aggregate utility functions has met insurmountable technical and conceptual problems; Sen (1988) reviews work in this area.

For such reasons among others, we are skeptical of loss function analysis (Wachter, 1991; Freedman, Wachter, Cutler and Klein, 1994). Now, BR claim (Section 2.2), our position is self-contradictory. On the one hand, we believe "that no index for summarizing the evidence from data is an unambiguous measure of whether one estimate is better than another." On the other hand, we think census adjustment is an empirical issue. But where is the contradiction? Their argument depends on an implicit assumption: the only way to make an empirical assessment is to pick a loss function and run the numbers. They cannot be right. For one thing, optimal estimators—optimal by all sensible criteria—seldom exist. That would seem to be the central lesson from 50 years of work on estimation theory. Indeed, a few pages later, BR concede that "it is impossible to determine a single loss function that is appropriate for evaluating every effect of an error in census numbers" (Section 2.6, item 1).

No single loss function can do the job, for reasons given above: loss functions do not measure the real social costs of errors, and different summaries of error distributions may give the advantage to different estimators. Even more to the point, when risks (expected losses) have to be estimated, different statistical assumptions about the data may give radically different conclusions. At best, loss function analysis is only part of an empirical assessment. Indeed, BR's bottom line (at the end of Section 2.2) is quite modest: they "see no reason to ignore the evidence from research on loss function analysis." Neither do we. Section 5 below reviews this evidence, as it bears on the key issue—the accuracy of state population shares.

We turn to the narrower question in our paper: does heterogeneity create bias in estimated risks? Our Table 9 covered seven proxies and showed that the biases could be large or small, proadjustment or antiadjustment. We conclude that the data cannot decide the issue. There is a critical parameter that does not seem estimable: the correlation between errors in the adjusted shares and the adjustments themselves.

Belin and Rolph (Section 4) make two arguments on this topic: (i) on average, over the seven proxies in the table, the results "favor neither adjusted nor unadjusted figures" and (ii) this is to be expected on theoretical grounds because "it does not seem obvious why such a correlation [between errors and adjustments] would occur." The structure of these arguments is illuminating. In (i), lack of knowledge about which proxy best represents the undercount is replaced by a model for ignorance (the uniform distribution); then strong conclusions are drawn from the model. In (ii), an unknown correlation is replaced by 0. The unknown is made known by models, whose conclusions are to be accepted unless they can be disproved.

Belin and Rolph describe our position as "adversarial," preferring "the more balanced interpretation in Fay and Thompson (1993)"; EFK appeal to the same paper for the same reason. But what do Fay and Thompson say? "Failure of the homogeneity as-

sumption is potentially a larger source of error than all errors explicitly included in the total error model and loss function analysis"; Fay and Thompson go on to ask, "what sense can then be made of the loss function analysis?" After reviewing the bureau's research program, they conclude:

> A much larger base of experience along the same lines may suggest principles or test procedures to distinguish the circumstance under which the loss function analysis is robust [against heterogeneity] compared to instances leading to its breakdown .... In 1990, the issue of heterogeneity affected the most constitutionally important statistics: the population of the states ... future designs should set realistic and clearly defined reliability goals for direct estimates [not using the synthetic assumption] for states. (Fay and Thompson, 1993, pages 81–83)

Will EFK and BR accept that formulation?

Imputation Models

The 1991 estimated undercount from the PES was about 5.3 million persons (net, nationwide). However, 4.1 million persons (weighted to national totals) were "unresolved" in the P-sample: it could not be determined from the PES fieldwork whether or not these cases matched to the census. Their match status was imputed by—you guessed it—a model. This imputation model has a powerful effect on estimated adjustments at the state level (Wachter, 1991), and the model has two very peculiar features (Wachter, 1993b). To explain these, we refer to the Evaluation Followup Survey (the EFU), a second survey that tried much later to reinterview a sample of the unresolved P-sample cases.

(i) Nearly 32% of the unresolved P-sample cases fell into a special class, which we shall call the Q-class (Q for "question marks"). This Q-class consisted of cases about whom only minimal information was obtained in the first wave of PES interviews. It was not judged cost-effective to send such cases to PES follow-up, still less to the Evaluation Followup many months later. Instead, match status was imputed by assuming that cases in the Q-class were like PES respondents with the strongest information, namely, those who could be matched by the computer after the first wave of PES interviews. This is a peculiar assumption.

(ii) The remaining 100% – 32% = 68% of the unresolved PES cases were in the EFU sampling frame; fieldwork was done by the EFU on a sample of these cases, to validate the imputation model. However, the EFU could resolve only 41% of the cases that were sent to EFU. Thus, 59% remained unresolved (match status to the census remained indeterminate). The unresolved group must consist of cases with weak data. However, the imputation model says that these weak cases match to the census at a much higher rate than the cases resolved in EFU—the cases with strong data; details are in Appendix 1. This is equally peculiar.

In Sections 2.3 and 5.7, BR now try to defend the imputation model, using data from the EFU. However, the EFU resolved only $(1 - 0.32) \times 0.41 = 28\%$ of the unresolved PES cases. What about the remaining 72% of the PES imputations? Were these right or wrong? The EFU results cannot tell us, because the EFU could not decide the match status of those persons. This difficulty has been explained in the journals (Wachter, 1993b) and in private correspondence, a snippet of which BR reproduce. BR respond by calling our position "very extreme," but where do they draw the line? If 90% of the data are missing? 95%? 99%? Or do they think that any amount of missing data can be filled in, just by making up models?

The model's assumptions are silly. It is even sillier to claim these assumptions have been validated by the EFU, when the groups in question were either not sent to EFU or remained unresolved in EFU.

When Leo Breiman looks at the EFU data—all the EFU data—he finds another paradox: the rate of unresolved in EFU goes up with predicted match rate. The Census Bureau did the same analysis (Gbur, 1991c). Belin and Rolph say that our friend Leo is "play[ing] games with ... percentages" by looking at all the data: BR insist that percentages should be based only on cases resolved in EFU. That is because BR look at the data only through the prism of their models. BR are playing games with models; Leo is blowing the whistle.

Some Points of Detail

We think BR are wrong on many points of factual detail, and their interpretations of the scientific literature are often quite strained. We give three examples.

Documentation. According to BR (Section 2.3), the adjustment process should have been, and was, "fully documented with its assumptions spelled out." We have replicated many parts of the Bureau's smoothing model and loss function calculations (Freedman et al., 1993, page 416; Freedman, Wachter, Cutler and Klein, 1994, page 277). However, we had a lot of friendly help from Bureau personnel. An investigator who seeks to do such work just on the basis of the printed record will have quite a frustrating time: the

documentation is maddeningly cryptic about many critical issues.

Presmoothing. Before using estimated variances to smooth the adjustment factors, the Bureau "presmooths" the variances: indeed, smoothing without presmoothing would have cut the estimated undercount from 2.1% to 1.2% (Ericksen and Tukey, 1991, page 2). Our opponents have defended presmoothing on technical grounds, but we believe their arguments are essentially circular (Freedman et al., 1993, pages 383–385). In response, BR assert (at the end of Section 3.8) that the Census Bureau decided "to carry out presmoothing for bias-reduction reasons" around 1988 or 1989. They are mistaken. Some form of presmoothing was indeed under active consideration by the Bureau, at least since the 1988 test census in St. Louis (the "dress rehearsal"). However, the key decisions—whether to presmooth and how to do it—were still being debated in 1991. For example, one of us participated in such discussions with senior personnel from the Bureau and the Special Advisory Panel on May 16 of that year.

Hindsight is 20/30. There was an adjustment to the census proposed in 1991 and evaluated in 1991, based on data available in 1991, including data on estimated errors in the PES. Subsequently, additional errors were discovered (Section 6 below). Some of these were corrected, leading to another adjustment proposed in 1992 for the intercensals, and evaluated based on data available in 1992 about the remaining errors.

In Section 5.2, BR suggest that evaluating the 1991 adjustment based on the 1992 error estimates (rather than the 1991 error estimates) makes the case for the 1991 adjustment even stronger. That just misreads the literature. Shown below are three possible evaluations:

(a) the 1991 adjustment evaluated using 1991 error estimates;
(b) the 1992 adjustment evaluated using 1992 error estimates;
(c) the 1991 adjustment evaluated using 1992 error estimates.

The papers cited by BR focus on (a) and (b). Some of the work for (c) has been done; but BR ignore the results, which are summarized below.

5. LOSS FUNCTION ANALYSIS

We now consider loss function analysis for the 1991 proposed adjustment to the 1990 census; 1,392 poststrata are in full sway. The focus is on state population shares; "loss" is squared error in state shares (Woltman et al., 1991; Mulry and Spencer, 1993). Let G be the Census Bureau's estimated covariance matrix for the adjustments, as derived from their smoothing model; let H be the estimated covariance matrix for estimated biases in those adjustments. (Although H does not affect point estimates of risk, it does come into the estimated standard errors.)

As noted earlier, G is biased downward, by a factor of 2 or 3. It is shown in Freedman, Wachter,Cutler and Klein (1994, pages 268ff) that H too is biased downward, by a factor on the order of 50 or 100. That is a remarkable claim, and we stand behind it. In brief, the Bureau estimated biases on the basis of an Evaluation Follow-Up sample that was perhaps 7% of the size of the PES (in terms of households, at any rate). The Bureau is claiming variances about 6 times smaller than raw variances in the PES, rather than $1/0.07 = 14$ times larger: $6 \times 14 = 84$ is a big factor. The bureau achieved its reduction in apparent variance by computing H on the basis of an allocation scheme, omitting any uncertainty due to variation in error rates across poststrata or geography. (For a tantalizing hint on the existence of this difficulty, see Fay and Thompson, 1993, page 79.)

What are the implications for loss function analysis? The first four lines of Table R2 allocate bias from evaluation strata to individual poststrata using the bureau's "PRODSE" method. The last four lines of the table increase the level of bias to 50% of the undercount (Section 6) and allocate in proportion to the undercount. The effects of correcting G and H are shown too. For details, see Freedman, Wachter, Cutler and Klein (1994).

With the Bureau's way of doing things, reported in the first line of the table, the estimated risk difference (census risk − adjustment risk) is 667 parts per 100 million, with an SE of 281: adjustment is a winner. In the last line of the table, which strikes us as the most realistic, the census comes out ahead; the difference is not significant. With intermediate lines, the case for adjustment is hardly convincing. In sum, loss function analysis is driven by the models that underlie it not by the data.

Line 1 in Table R2 may also be contrasted with line 5, where bias is allocated as 25% of the undercount. The contrast demonstrates that the allocation scheme determines the outcome even if we grant the Bureau their estimates for G, H and overall level of bias. (The Bureau's 1991 loss function analysis included an allowance for bias amounting to nearly 25% of the net undercount; Mulry, 1991, Table 14.)

Table R2 is computed on the basis of the synthetic assumption. If undercount rates are heterogeneous within poststrata across states, that would be another source of bias in the estimated risk differences. How large is the effect? We do not know, and no-

TABLE R2
Impact of allocation schemes for state-level biases, correction of final variances and correction of variances in estimated biases: Estimated risk difference, census risk − adjustment risk, and standard error; units are parts per 100 *million*

Allocation of bias	Correction factor for G	Correction factor for H	Estimated risk difference	SE
PRODSE	1	1	667	281
PRODSE	1	50	667	890
PRODSE	2	1	542	371
PRODSE	2	50	542	885
0.25 × undercount	1	1	193	199
0.50 × undercount	1	1	−125	156
0.50 × undercount	1	50	−125	859
0.50 × undercount	2	1	−250	169
0.50 × undercount	2	50	−250	821

body else does either. For example, if ALL is a good proxy for undercounts and we can extrapolate from 357 poststrata to 1,392, the bias in estimated risks will be small. If DIFF is the better analog, the bias is of the same order of magnitude as the estimated risk difference itself, and favors adjustment.

6. BIAS IN THE PES

We distinguish between "measured" and "unmeasured" bias in the PES. Measured bias is caused by matching error, census day address error and so forth. In principle, such biases can be estimated by reinterviewing and rematching studies, although the difficulties are numerous. Generally, the measured biases cause the PES population estimates to be too high. Correlation bias, on the other hand, is unmeasured. Typically, this bias occurs when (even within a poststratum) people who are missed by the census are also more likely to be missed by the PES. This sort of bias makes the PES estimates too low. There are "unreached people," missed both by the census and by the PES adjustment.

We discuss the measured biases first, then return to correlation bias. In July 1991, the estimate for the net national undercount was 2.1%, with measured biases thought to total 0.7 percentage points. Additional errors were discovered in the PES (BR, Section 5.2; Fay and Thompson, 1993, page 74; Bureau of the Census, 1993, page 75.) These errors reduced the undercount estimate by 0.5 percentage points. The measured biases were still thought to total 0.7 percentage points, leaving $2.1 - 0.5 - 0.7 = 0.9\%$ for the undercount. In other words, on the bureau's latest figures, measured biases amounted to 57% of the 1991 estimated net undercount: $(0.5 + 0.7)/2.1 = 0.57$. On Breiman's figures (Section 6.1), measured biases amount to 80% of that estimated undercount. Either way, most of the 1991 estimate represents bad data rather than undercount.

Correlation Bias

Correlation bias cannot be measured directly when the census and the system designed to correct the census both miss the same people. There can be no direct evidence about people if the surveys cannot find them. Thus, correlation bias is measured indirectly, using demographic analysis to make a second estimate of the national population (but see, e.g., Darroch, Fienberg, Glonek and Junker, 1993).

Commenting on this rather obvious point, Breiman says that estimates of correlation bias have only a "tenuous connection with any data." Belin and Rolph respond (Section 5.5):

> Correlation-bias estimates have more than a "tenuous connection with any data"; on the contrary, estimates are based on combining PES data with evidence from vital records about the ratio of the size of the male population to that of the female population. These data are used to estimate a parameter that characterizes dependence in omission rates between the census and PES (Bell, 1993).

We defer to Fay and Thompson (1993, page 76), who seem to agree with Breiman:

> "An important component of the total error model, correlation bias, could not be directly measured at any level, and was only indirectly inferred nationally by use of sex ratios derived from demographic analysis. Distribution of the national results relied entirely upon models that could not be checked against any direct evidence. . . ."

Numerical results from Bell's model give some insight into plausibility of assumptions. At the national level, the model says:

1. the PES estimates missed 890,000 white males total, of whom 13 (this is not a typo) are in the prime age group between 20 and 30;
2. the PES estimates missed 760,000 black males, of whom −28,000 are under age 10.

These results are incredible. Indeed, a direct comparison of demographic analysis with the PES estimates shows the latter found 190,000 too many white males, not 890,000 too few.

Likewise, when EFK say that Mulry and Spencer "used their best estimate of correlation bias," EFK mean that Mulry (1991) took Bell's national totals and disaggregated them to the 13 evaluation strata

and the 1,392 poststrata. (It is the latter numbers that feed into the loss function analysis, through the "total error model.")

The rationale for Mulry's algorithm was not provided. EFK and BR may not have examined the calculations; if they do, they will find that:

1. Mulry uses Bell's numbers as if they stood for correlation bias alone when in fact they are estimates of net error (correlation bias offset by the measured biases).
2. Mulry's detailed numbers do not add up to Bell's totals. Bell starts out with 1,652,000 men and no women; Mulry ends up with 285,000 men and 307,000 women: massive numbers of men disappear or are converted into women.

(These figures are not reported by Mulry, but can be estimated with reasonable precision from the available data.) Fay and Thompson (1993, page 76) sum up as follows:

> Unfortunately, the integration of [Bell's] results into the total error model was highly problematic .... Thus, it is impossible satisfactorily to characterize by any explicit model the nature of the realized estimates incorporated into the total error model.

To defend the Census Bureau's model for correlation bias, EFK and BR brush aside these facts, blurring the distinction between measurements and allocation schemes. Correlation bias was not measured, and the estimates have only a tenuous connection with reality, just as Breiman said.

To say a bias is unmeasured is different from saying that it is zero, although EFK pretend not to understand that distinction either. Despite wide uncertainties in all the figures, direct comparisons between demographic analysis and the PES, with measured biases taken into account, suggest that correlation bias is substantial (Wachter, 1991). At the national level, large numbers of black males seem to be missing from the adjusted census counts. What geographical areas are they missing from? No one knows. A reasonable opinion is that they are missing from inner cities with large minority populations, in the northeast and midwest.

Had the census been adjusted, the state population share of New York, for example, would have gone *down*: down, not up. Why did the 1990 adjustment bring down the population shares of New York and other such states? Correlation bias is a prime suspect. The bureau's total error model and loss function analysis cannot detect salient errors in the proposed adjustment to state shares, because the bureau's stylized rules for allocating unreached people to geography have little connection with facts on the ground.

7. DISCUSSION

Would the 1991 adjustment have improved on the census? Proponents of adjustment give strongly positive answers to this question in the academic literature, in the administrative record, and in the courtroom. For example, Ericksen (1991, page 3) writes:

> The only reasonable conclusion is that the adjusted count is more accurate than the unadjusted count...under any reasonable basis of comparison, the PES-adjusted enumeration is more accurate than the unadjusted census enumeration. Those adjusted results have also been shown to be robust to variations in reasonable alternatives to the PES "production procedures," and to variations in the statistical models used to generate the adjusted figures.

According to Fienberg (1992a, page 28), "the results of the Census Bureau's evaluation studies clearly supported the use of adjustment for the 1990 census results." Here is Rolph (1993, page 97), summarizing the evidence on loss function analysis, as presented by him and other plaintiffs' witnesses in *New York v. Department of Commerce*: "the Bureau's analysis clearly demonstrated that the adjusted counts were an unambiguous improvement on the original enumeration." These are sweeping claims, especially when compared to the results in Breiman or in Table R2 above.

The first line of Table R2, replicating the Census Bureau's analysis for the 1991 adjustment, is worth another comment. The 667 is the difference between an estimated risk (squared error) of 734 for the census, and 67 for adjustment; units are parts per 100 million. Over the 50 states and Washington, D.C., the r.m.s. error in population shares from the census is estimated as $\sqrt{734 \times 10^{-8}/51} = 0.04\%$. (The error distribution is quite skewed, which creates additional complications; Freedman, Wachter, Cutler and Klein, 1994, pages 255–259.) Our opponents contend that these errors can be reduced by an order of magnitude in size, if only we would agree to use their adjustment technology. Given the scale of the errors, that is—or should be—an astonishing claim.

EFK invent for us the position that the census is perfect and that only perfect models are usable (Fienberg, 1992a, page 27); BR have us assuming that undercount rates are constant across poststrata (Section 2.5). Thus, instead of showing that their models are accurate enough to correct miniscule errors in census shares, they remind us that the world is imperfect. That sets up their favorite rhetorical trick, which they play over and over again: justifying their assumptions as being less imperfect than the ones they have created for us.

The imperfection of the world is an argument much loved by modelers. The work must be done; the lesser evil must be chosen; and the best is the enemy of the good (Fienberg, 1992a, page 27; BR, Section 2.3). This argument has little force in the present context, when "the work that must be done" is only the creation of a technical record to defend a prior set of models.

Listen to them. They can adjust Stockton, with almost no data from that city, just by making the right assumptions (that demography overrides geography). They can validate their imputation model, when 75% of the validation data are missing, just by making the right assumptions: cases with weak data are easier to match than cases with strong data. They can measure correlation bias in New York, just by making the right assumptions, although they cannot quite explain what those assumptions are.

The best we can say for adjustment is that (i) it can fix the estimated differential undercounts at the national level and (ii) its impact on the accuracy of state population shares cannot be determined with any great confidence. In our opinion, however, adjustment is likely to degrade the accuracy of the state shares, and for substate areas adjustment seems even worse.

What will happen in the next census? Unless the analytical mistakes of 1990 can be recognized, they are likely to be repeated on an even larger scale in the year 2000. Efforts by BR and EFK to defend the mistakes of the past may cast a long shadow.

APPENDIX 1: VALIDATING THE IMPUTATION MODEL WITH EFU DATA

We document our calculations for missing data in the EFU as follows. Gbur (1991c, Table 3.1) shows 41% resolved in EFU out of the 2.8 million unresolved-in-P-sample cases sent to EFU, weighted to national totals. There were a total of 4.1 million unresolved P-sample cases in the bureau's "Advisory Use File": $(4.1 - 2.8)/4.1 = 0.32$, that is, 32% of the unresolved P-sample cases fell into the "Q-class, page 5 above." Our discussion focuses on the unresolved cases in the P-sample; other issues would arise for the E-sample (Breiman, 1994, Section 7).

Our next object is to state the issue in dispute between Breiman and BR (Section 5.7). Some notation will be helpful, although the argument will be informal. Let M be match status in the census (1 is a match, 0 is a nonmatch). Let R indicate inclusion/exclusion and resolved/unresolved in EFU: -1 is excluded, 0 is included but unresolved, 1 is resolved; the Q-class corresponds to $R = -1$. Let Z be the covariates in the imputation model that BR are defending; let $\hat{p}$ be the match probability imputed by the model; $\hat{p}$ is a function of Z, and Z is computed from PES data. For cases unresolved in the PES, match status is unknown; for a sample of these cases, M is determined by the EFU, but not for cases excluded from the EFU sampling frame or cases left unresolved in the EFU.

To validate their model, BR need to show that

$$P\{M = 1 \mid Z\} \approx \hat{p}. \tag{1}$$

Their claim is weaker:

$$\textit{BR's claim } P\{M = 1\} \approx E\{\hat{p}\}, \tag{2}$$

where "P" stands for weighted relative frequencies and "E" for weighted averages. What the EFU data show is weaker yet:

$$P\{M = 1 \mid R = 1\} \approx E\{\hat{p} \mid R = 1\} \approx 0.32. \tag{3}$$

It is this agreement that BR emphasize so strongly in Section 2.3. (There is an irritating numerical coincidence: 0.32 is also the fraction of Q-class cases.) To get (2) from (3), BR need

$$P\{M = 1 \mid R < 1\} \approx E\{\hat{p} \mid R < 1\}. \tag{4}$$

Of course, $P\{M = 1 \mid R < 1\}$ is not known: that is one implication of $R < 1$. You might think the trail ends here, but BR do not give up.

The next calculation is slightly indirect, because there are gaps in the documentation. As usual,

$$\begin{aligned} E\{\hat{p}\} &= E\{\hat{p} \mid R = 1\} \times P\{R = 1\} \\ &\quad + E\{\hat{p} \mid R < 1\} \times P\{R < 1\}. \end{aligned} \tag{5}$$

We know from the Advisory Use File that $E\{\hat{p}\} = 0.53$, and $P\{R = 1\} = 0.41 \times 2.8/4.1 \approx 0.28$, as before. By (3), $E\{\hat{p} \mid R = 1\} \approx 0.32$. Thus,

$$\begin{aligned} E\{\hat{p} \mid R < 1\} &= \frac{E\{\hat{p}\} - E\{\hat{p} \mid R = 1\} \times P\{R = 1\}}{P\{R < 1\}} \\ &\approx \frac{0.53 - 0.32 \times 0.28}{1 - 0.28} \\ &\approx 0.61. \end{aligned} \tag{6}$$

That $E\{\hat{p} \mid R < 1\} \approx 0.61$ is an empirical fact. Via (4), BR's claim (2) entails $P\{M = 1 \mid R < 1\} \approx 0.61$. By (3), $P\{M = 1 \mid R = 1\} \approx 0.32$; this is another empirical fact. Thus, BR's claim (2), coupled with the data, entails

$$P\{M = 1 \mid R < 1\} \approx 2 \times P\{M = 1 \mid R = 1\}. \tag{7}$$

You might think, as we do, that $P\{M = 1 \mid R < 1\}$ should be substantially lower than $P\{M = 1 \mid R = 1\}$.

But it must be nearly twice as high if you accept BR's claim (2). By their account, the cases that were excluded from EFU or were unresolved in EFU are much more likely than the resolved cases to match to the census. Belin et al. (1993, page 1164) actually say this, after presenting a calculation similar to ours. There is simply no evidence to support their bizarre story: M is unknown for $R < 1$. (Belin et al.'s arithmetic is wrong because they include $R = -1$ in one place and exclude it in another; even if they had gotten the arithmetic right, however, they would still be in the same logical mess.)

So far, the discussion has been about all the P-sample imputations, those excluded from EFU and those included in EFU. Breiman's paradox is only about cases included in EFU:

(8) $$P\{R = 1 \mid \hat{p} \text{ and } R > -1\} \text{ decreases as } \hat{p} \text{ increases.}$$

To explain (8), BR considered a group of cases with weak data, whose "names are not recorded." Let $W = 1$ for cases in this group, and $W = 0$ for other cases. Belin and Rolph have argued (personal communication) that

(9) $$E\{\hat{p} \mid W = 1\} \text{ is high}$$

and

(10) $$P\{R = 0 \mid W = 1\} \text{ is high.}$$

If (9) and (10) were right, they could explain Breiman's paradox:

(11) the weak cases do not get resolved in EFU, and that *lowers* the match rate among cases with big $\hat{p}$'s.

We consider these steps in turn. Assertion (9) implies (for the EFU universe) that unresolved P-sample cases with weak data match to the census at higher rates than other cases. Our view, of course, is that PES forms have weak data because there is only weak evidence that the corresponding people exist in the first place. Then cases with weak data will match to the census at lower rates, and (9) seems questionable at best. Assertion (10) seems right in general, but problematic for BR's chosen group of weak cases (see below). If you grant (9) and (10), then (11) is a good argument.

None of this can have much numerical impact: the group of weak cases considered by BR (whose "names are not recorded") was, with minor exceptions, not sent to EFU at all. These are cases with $R = -1$, not $R = 0$. They have nothing to do with Breiman's finding. The only additional argument made by BR is that they "wholeheartedly disagree" with us (personal communication). Their hearts are in the right place. Now, can we appeal to their heads?

APPENDIX 2: LEGAL PROCEEDINGS

In 1988, New York City (among others) filed suit in federal district court to compel the Department of Commerce to adjust the census. In 1992, the court ruled against adjustment, on fairly narrow grounds (Freedman, 1993b, page 106). New York went to an appeals court, the "Second Circuit," which vacated the judgment of the district court and remanded for further proceedings. The Supreme Court may yet determine the issue, because the Department of Commerce prevailed in the Sixth and Seventh Circuits: *Detroit v. Franklin* and *Chicago v. Department of Commerce*. Legal commentaries by statisticians should be taken with several grains of salt—and a dash of vinegar—but here we go.

Generally, an appeals court will rely on the district court's findings of fact. The Second Circuit held that adjustment was more accurate than the census, paraphrasing the district court to say that "for most purposes and for most of the population. . .adjustment would result in a more accurate count than the original census" (page 37 of the Second Circuit's typed opinion, dated August 8, 1994). However, the Second Circuit seems to have picked and chosen among the findings of the district court. For example, according to the district court, plaintiffs had failed to "illustrate affirmatively the superior accuracy of the adjusted counts [at the state and local level] for any reasonable definition of accuracy" (*Federal Supplement* **822** 924). The Second Circuit simply ignored the findings that did not suit.

The Second Circuit emphasized numeric accuracy and criticized the Secretary of Commerce for giving priority to "distributive accuracy," that is, accuracy of population shares for geographic areas (Second Circuit, page 42). Curiously enough, the Second Circuit based its own legal argument on cases (including *Baker v. Carr*, *Wesberry v. Sanders*, *Reynolds v. Sims* and *Karcher v. Daggett*) that deal with distributive accuracy (Second Circuit, pages 29ff). There is some difficulty, discussed briefly by the Second Circuit on pages 38ff, in applying these cases to the federal government; however, such legal issues are grist for another article by different authors.

Coming back to statistics, the Second Circuit's premise seemed to be that minority groups with large undercounts at the national level are concentrated in states whose shares would be adjusted upward. This is a fallacy. The same area often contains members of relatively overcounted groups along with mem-

bers of relatively undercounted groups. Therefore, an area's share often goes down as a result of adjustment, not up, despite a concentration of minorities. Urban blacks have an undercount three times that of the rest of the population, according to the PES; but 55% of them live in states that would lose population share if the adjustment were implemented.

The state with the largest number of blacks in the country, New York, would have its share adjusted downward. Pennsylvania, with nearly a million blacks, would lose a seat in Congress. The impact of adjustment on the constitutional right to equal representation has to be assessed area by area, in terms of demographic makeup and proposed adjustment factors. The Second Circuit did not come to grips with the statistical facts.

ADDITIONAL REFERENCES

AUSTRALIAN BUREAU OF STATISTICS (1990). *Census 86: Data Quality Undercount*, Catalogue No. 2607.0.

BELL, P., CORNISH, J., EVANS, J. and VINCENTE, W. (1993). Small area estimation of census undercount. Paper presented at the 49th Session of the International Statistical Institute.

BIEMER, P. and FORSMAN G. (1992). On the quality of reinterview data with application to the current population survey. *J. Amer. Statist. Assoc.* **87** 915–923.

BUREAU OF THE CENSUS (1975). Coverage of population in the 1970 census and some implications for public programs. Current Population Reports, Series P-23, No. 56, Government Printing Office, Washington, D.C.

CHOI, C. Y., STEEL, D. G. and SKINNER, T. (1988). Adjusting the 1986 Australian census count for under-enumeration. *Survey Methodology* **14** 173–189.

DIAMOND, I. (1993). Where and who are the missing million? In *Regional and Local Statistics: Statistics Users Council Conference.* 132–145. IMAC Research, Esher.

ERICKSEN, E. (1991). Letter report to Secretary Mosbacher, dated June 21/91. Reproduced in Department of Commerce (1991b). Office of the Secretary. Decision on whether or not a statistical adjustment of the 1990 decennial census of population should be made for coverage deficiencies resulting in an overcount or undercount of the population; explanation. Report dated July 15, Washington, D.C. Vol. II. Also see *Federal Register* **56** 33,582–33,642 (July 22).

ERICKSEN, E. and TUKEY, J. (1991). Letter report to Secretary Mosbacher, dated July 11/91. [Reproduced in Department of Commerce (1991b). Office of the Secretary. Decision on whether or not a statistical adjustment of the 1990 decennial census of population should be made for coverage deficiencies resulting in an overcount or undercount of the population; explanation. Report dated July 15, Washington, D.C. Vol. III. Also see *Federal Register* **56** 33,582–33,642 (July 22).

FIENBERG, S. E. (1994). An adjusted census in 1990? Trial judgment set aside *Chance* **7** (4) 31-32.

HIMES, P. and CLOGG, C. (1992). An overview of demographic analysis as a method for evaluating census coverage in the United States. *Population Index* **58** 587–612.

KADANE, J. B., MEYER, M. M. and TUKEY, J. W. (1992). Correlation bias in the presence of stratum heterogeneity. Technical Report 549, Dept. Statistics, Carnegie Mellon Univ.

KIM, J. (1991). 1990 PES evaluation project P12: evaluation of synthetic assumption. Technical report, Bureau of the Census, Washington, D.C.

MULRY, M. (1991). 1990 Post Enumeration Survey evaluation project P16. Total error in PES estimates for evaluation poststrata. Technical report, Bureau of the Census, Washington, D.C.

OFFICE OF POPULATION CENSUSES AND SURVEYS (1993). How complete was the 1991 Census? *Population Trends* **71** 22–25.

ROBINSON, J. G. (1994). Use of analytic methods for coverage evaluation in the 2000 census. Paper presented at Population Association of America Meetings.

SEN, A. (1988). Social choice. In *The New Palgrave: A Dictionary of Economics* **4** 382–393. MacMillan, London.

SKINNER, C. J. (1991). The use of synthetic estimation techniques to produce small area estimates. Social Survey Division New Methodology Series NM18, Office of Population Censuses and Surveys, London, U.K.

STEEL, D. and POULTON, J. (1988). Geographic estimates of under-enumeration. In *Proceedings of the Survey Research Methods Section* 119–128. Amer. Statist. Assoc., Alexandria, VA.

TRICKETT, P. (1992). Preliminary assessment of the 1991 census undercount. Paper presented to the Sixth National Conference of the Australian Population Association.

WACHTER, K. (1991). Recommendations on 1990 census adjustment, dated June 17/91. Reproduced in Department of Commerce (1991e). Office of the Secretary. Decision on whether or not a statistical adjustment of the 1990 decennial census of population should be made for coverage deficiencies resulting in an overcount or undercount of the population; explanation. Report dated July 15, Washington, D.C. Vol. II. Also see *Federal Register* **56** 33,582–33,642 (July 22).

WACHTER, K. (1994). Discussion of "Use of analytic methods for coverage evaluation in the 2000 census," by J. G. Robinson, Population Association of America Meetings. Unpublished.