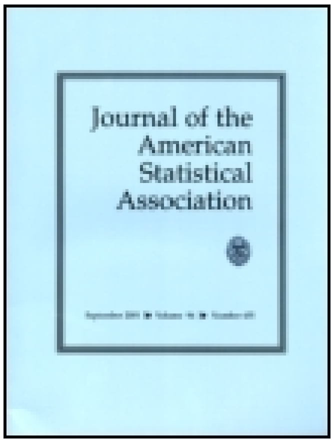

Journal of the American Statistical Association

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Model-Based Direct Adjustment

Paul R. Rosenbaum [a]

[a] Department of Statistics , Wharton School, University of Pennsylvania , Philadelphia , PA , 19104-6302 , USA

Published online: 12 Mar 2012.

Model-Based Direct Adjustment

PAUL R. ROSENBAUM*

Direct adjustment or standardization applies population weights to subclass means in an effort to estimate population quantities from a sample that is not representative of the population. Direct adjustment has several attractive features, but when there are many subclasses it can attach large weights to small quantities of data, often in a fairly erratic manner. In the extreme, direct adjustment can attach infinite weight to nonexistent data, a noticeable inconvenience in practice. This article proposes a method of model-based direct adjustment that preserves the attractive features of conventional direct adjustment while stabilizing the weights attached to small subclasses. The sample mean and conventional direct adjustment are both special cases of model-based direct adjustment under two different extreme models for the subclass-specific selection probabilities. The discussion of this method provides some insights into the behavior of true and estimated propensity scores: the estimated scores are better than the true ones for almost the same reason that direct adjustment can outperform the sample mean in a simple random sample. The method is applied to a nonrandom sample in an effort to estimate a discrete distribution of essay scores in the College Board's Advanced Placement Examination in Biology.

KEY WORDS: Propensity scores; Poststratification; Standardization; Conditional inference.

1. INTRODUCTION: THE NEED FOR ADJUSTMENT; NOTATION; DIRECT ADJUSTMENT; IGNORABLE SELECTION

1.1 Introduction

In observational studies, in survey nonresponse, in the equating and norming of psychological tests, and in other applications, adjustments are used in an effort to estimate population means, moments, and distributions from a sample that is not representative of the population. A conventional method of adjustment divides the sample into strata or subclasses and reweights subclass means by using population frequencies. This method is known variously as poststratification (Holt and Smith 1979; Little 1982; Oh and Scheuren 1983), subclassification (Cochran 1968; Rubin 1977), or direct standardization (Fleiss 1981, sec. 14.4; Mosteller and Tukey 1977, sec. 11). The purpose of this article is to develop a number of technical improvements in the method of direct adjustment.

This section briefly introduces a motivating example, defines some notation, discusses conventional direct adjustment, its strengths and limitations, and finally discusses a condition—ignorable selection—without which neither conventional adjustments nor the proposed methods will yield appropriate estimates. Section 2 discusses some elementary theoretical issues; these are summarized in the introduction to that section. Section 3 returns to the example, applying model-based direct adjustment and examining its performance.

1.2 An Example of the Need for Adjustments: The College Board's 1982 Advanced Placement Exam in Biology

The College Board's 1982 Advanced Placement Examination in Biology included (a) 120 multiple-choice questions or items in three groups of 40 items testing cellular and molecular biology, organismal biology, and population biology, and (b) six free-response or essay questions grouped into three pairs, with each student answering his own choice of one essay within each pair.

Students selecting one essay rather than another may differ systematically, as was the case with the essay pair consisting of Essay 5 and Essay 6. In particular, the mean number of the multiple-choice items answered correctly was 70.66 for the 4,129 examinees selecting Essay 5 and 65.23 for the 11,547 examinees selecting Essay 6, with a two-sample t statistic of 16.7. Figure 1 is a stem-and-leaf display (Tukey 1977) of 120 standardized differences in proportions of examinees answering correctly each of the 120 multiple-choice items. For 118/120 items, a higher proportion of correct responses is observed among examinees selecting Essay 5. For two items, a higher proportion of correct responses is observed among examinees selecting Essay 6: one difference (Item 95) is negligible, the other (Item 4) is substantial. Both Item 4 and Essay 6 ask about animal behavior, possibly indicating that students who knew more about animal behavior selected the essay on that topic. Closer examination of the pattern of item responses in this example and in several other similar examples suggests that examinees selecting different essays differ in fairly subtle ways; in particular, they differ in ways not reflected in their total scores (Rosenbaum 1985).

For a variety of purposes associated with test analysis and test scoring (e.g., Coffman 1971, pp. 289–290), it is of interest to estimate the distributions of essay scores that would have been obtained had all students written the same essay, either Essay 5 or Essay 6. Each essay is scored on a 15-point scale. Some form of adjustment is required, since the examinees selecting one essay instead of the other are not representative of the combined group of all of the examinees. An estimate of the Essay 5 distribution for all examinees will be obtained by grouping examinees into about a thousand subclasses on the basis of the 40-item subscores and certain individual item responses. Unfortunately, with a thousand subclasses, conventional direct adjustment or standardization cannot be applied, because many subclasses contain no examinees who have selected Essay 5 (see Sec. 1.4 for elaboration). This article develops an alternative to conventional direct adjustment that han-

* Paul R. Rosenbaum is Associate Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6302. The author is grateful to P. W. Holland, S. Livingston, D. B. Rubin, an associate editor, and the editor (J. Kadane) for helpful comments or discussions, to C. Haag for permission to use the data in Section 3, and to B. Kaplan for assistance with computing.

Journal of the American Statistical Association
June 1987, Vol. 82, No. 398, Applications

		Depth
−7	9*	1
−6		
−5		
−4		
−3		
−2		
−1		
−0	4	2
0	45	4
1	14678889	12
2	01124577	20
3	333334569	29
4	00335555566778889	46
5	00011112222233333455666777889	(29)
6	02233445567899	45
7	02334446667	31
8	011244466899	20
9	02248	8
10	08	3
11	2	1

*Item 4

Figure 1. Stem-and-Leaf Display of Standard Differences for 120 Items for Essays 5 and 6 [i.e., $(p_{5j} - p_{6j})/\{[p_{5j}(1 - p_{5j})/n_5] + [p_{6j}(1 - p_{6j})/n_6]\}^{1/2}$, where for $j = 1, 2, \ldots, 120$, p_{5j} and p_{6j} are the proportions of students answering Item j correctly among those students who selected Essays 5 and 6, respectively, and n_5 and n_6 are the number of students who selected Essays 5 and 6]. Note that $n_5 = 4{,}129$, $n_6 = 11{,}547$, and 306 examinees did not write either essay. Source: Rosenbaum (1984d).

dles large numbers of subclasses and sparse data in a sensible and straightforward way.

1.3 Notation for a Finite Population With *S* Subclasses

A finite population of N individuals is divided into S strata or subclasses, numbered $s = 1, 2, \ldots, S$, where the number N_s of individuals in subclass s is known and $N = N_1 + N_2 + \cdots + N_S$. The ith individual in subclass s has a response $\mathbf{r}_{si}$ ($i = 1, 2, \ldots, N_s$), and we wish to estimate the population mean of the $\mathbf{r}_{si}$'s, namely

$$\bar{\mathbf{r}} = \frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{N_s}\mathbf{r}_{si} = \sum_{s=1}^{S}\frac{N_s}{N}\bar{\mathbf{r}}_s, \tag{1.1}$$

where $\bar{\mathbf{r}}_s = (1/N_s)\sum_{i=1}^{N_s}\mathbf{r}_{si}$ is the population mean response in subclass s. The vector $\mathbf{r}_{si}$ may contain cross-products or powers of some more basic set of variables, in which case $\bar{\mathbf{r}}$ contains higher moments of these basic variables. Write $\mathbf{R}$ for the matrix with N rows containing the $\mathbf{r}_{si}$'s, and write $\mathbf{X}$ for the N-dimensional vector of integers between 1 and S indicating the subclasses to which individuals belong. As is usually done in finite population sampling, $\mathbf{R}$ and $\mathbf{X}$ will be treated as fixed features of the finite population from which individuals are selected by a stochastic mechanism.

Responses are observed only for certain selected individuals. To distinguish the selected and unselected individuals, let $z_{si} = 1$ if $\mathbf{r}_{si}$ is observed, and let $z_{si} = 0$ if $\mathbf{r}_{si}$ is not observed. Let $\mathbf{Z}$ be the N-dimensional vector of z_{si}'s, let $n_s = \sum_{i=1}^{N_s} z_{si}$ be the number of selected units in subclass s, and let $n_+ = n_1 + n_2 + \cdots + n_S$ be the total number of selected units.

1.4 Direct Adjustment: Correcting for Disproportionate Selection by Assigning Population Weights to Subclass Means

When the population frequencies N_s ($s = 1, \ldots, S$) are known, an obvious estimator of $\bar{\mathbf{r}}$ is obtained by replacing the subclass specific population means—that is, the $\bar{\mathbf{r}}_s$'s in (1.1)—by the subclass specific sample means,

$$\bar{\bar{\mathbf{r}}}_s = \frac{\sum_{i=1}^{N_s} z_{si}\mathbf{r}_{si}}{n_s}, \tag{1.2}$$

thereby obtaining the estimator

$$\mathbf{D} = \sum_{s=1}^{S}\frac{N_s}{N}\bar{\bar{\mathbf{r}}}_s = \frac{1}{N}\sum_{s=1}^{S}\sum_{s=1}^{N_s}\frac{z_{si}\mathbf{r}_{si}}{(n_s/N_s)}. \tag{1.3}$$

Use of the estimator $\mathbf{D}$ is known variously as poststratification, subclassification, or direct standardization or adjustment. Note that in (1.3) the weight, N_s/n_s, assigned to an observed response $\mathbf{r}_{si}$ in subclass s is the reciprocal of the sample proportion of selected individuals in subclass s.

Direct adjustment has one unattractive property and two attractive ones. The first attraction is that direct adjustment does not require explicit modeling of the relationship between the response $\mathbf{r}_{si}$ and the variables used to construct the subclasses; this is particularly important in connection with educational tests and survey questionnaires in which $\mathbf{r}_{si}$ may contain several hundred item responses for each individual. The second attractive property is that, because direct adjustment reweights the observed responses, it produces parallel adjustments in the various coordinates of the response: as a result, directly adjusted covariance matrices are positive semidefinite and directly adjusted cumulative distribution functions are nondecreasing.

The principal difficulty with direct adjustment is that it can apply large weights to small quantities of data, often in a fairly erratic manner. In the extreme, n_s will equal zero for some subclasses, so $\bar{\bar{\mathbf{r}}}_s$ will be undefined, the weights N_s/n_s in those subclasses will be infinite, and $\mathbf{D}$ will be undefined.

Holt and Smith (1979) observed that, under simple random sampling of individuals, $\mathbf{D}$ is conditionally unbiased for $\bar{\mathbf{r}}$ given $\mathbf{n} = (n_1, n_2, \ldots, n_S)^T$, providing only that each n_s is greater than zero. In contrast the sample mean

$$\mathbf{M} = \frac{\sum_{s=1}^{S}\sum_{i=1}^{N_s} z_{si}\mathbf{r}_{si}}{\sum_{s=1}^{S} n_s}$$

is unbiased for $\bar{\mathbf{r}}$ under simple random sampling, but it is not *conditionally* unbiased for $\bar{\mathbf{r}}$. This distinction is important. As Holt and Smith noted, $\mathbf{M}$ is conditionally biased given $\mathbf{n}$; that is, $E(\mathbf{M} - \bar{\mathbf{r}} \mid \mathbf{n})$ is generally nonzero and varies with $\mathbf{n}$. Inspection of the sample counts $\mathbf{n}$ alone can reveal the direction of the conditional bias of $\mathbf{M}$ even before any response information is obtained (see Sec. 2.3). In more familiar though less explicit terms, **M** is affected by the between-subclass component of variability, which is to some extent predictable, whereas **D** is not so affected.

Direct adjustment is, of course, most often applied when individuals have not been selected by simple random sampling. Still, **D** is conditionally unbiased for $\bar{\mathbf{r}}$ under the weaker assumption of ignorable selection.

1.5 Ignorable Selection From a Finite Population

Selection is said to be *ignorable* when

$$\Pr\{\mathbf{Z} \mid \mathbf{R}, \mathbf{X}\} = \prod_{s=1}^{S} \prod_{i=1}^{N_s} e_s^{z_{si}}(1 - e_s)^{1-z_{si}} \qquad (1.4a)$$

$$\text{with } 0 < e_s \le 1 \quad \text{for} \quad s = 1, 2, \ldots, S, \qquad (1.4b)$$

where the e_s's may be unknown. In (1.4a), e_s is a conditional probability of selection given the fixed features of the population (**R**, **X**). Note in particular that when selection is ignorable, this conditional probability given (**R**, **X**) of selection for the ith individual in subclass s does not depend on his response $\mathbf{r}_{si}$. In addition, by (1.4b), under ignorable selection, every individual has a positive probability of selection. When selection is ignorable, the e_s's are propensity scores, as discussed in Rosenbaum and Rubin (1983a; 1984; 1985), Rubin (1983; 1984), Little (1984), and Rosenbaum (1984a).

It is both well known and easy to see that, when selection is ignorable, the conditional distribution of the selection indicators, **Z**, given (**n**, **R**, **X**) does not depend on the unknown e_s's, and assigns equal probability to each of the $\Pi_{s=1}^{S} \binom{N_s}{n_s}$ possible samples that select n_s individuals from subclass s, for $s = 1, 2, \ldots, S$—that is, the conditional distribution is that of a stratified random sample. It follows that under ignorable selection, **D** is conditionally unbiased for $\bar{\mathbf{r}}$ given (**n**, **R**, **S**) whenever **D** is defined, that is, whenever $n_s > 0$ for each s. Still, when the number, S, of subclasses is moderately large relative to N, some of the n_s's are likely to be zero and other are likely to be small. Elsewhere, in connection with hypothesis tests in observational studies, it has been observed that, under ignorable selection, when the e_s's satisfy a logit model, the conditional distribution of **Z** given the sufficient statistic for the parameter of the logit model does not depend on the unknown e_s's—that is, that we need not condition on all of **n** but can instead condition on a many–one function of **n** in eliminating the unknown e_s's (Rosenbaum 1984a). In the current context it is, therefore, natural to ask whether we may obtain an estimator that is defined when some $n_s = 0$, but which is nonetheless conditionally unbiased for $\bar{\mathbf{r}}$, providing we are willing to assume that the e_s's exhibit the kind of systematic variations permitted by a logit model. In fact, such an estimator does exist, and is discussed in Section 2.

Rubin (1976; 1978) first used the term ignorable in connection with Bayesian inference: there, selection is ignorable when factor of the likelihood that describes the stochastic selection process can simply be ignored in calculating the correct posterior distribution. A closely related condition, strong ignorability, was used by Rosenbaum and Rubin (1983a) for frequentist inference concerning superpopulation parameters. Rather than introduce a new term here, (1.4) will simply be called ignorable selection, since it differs from previous definitions only in minor details. Closely related ideas are discussed by Dawid (1976; 1979, sec. 7), Little (1982, sec. 3.4), Oh and Scheuren (1983), and Smith (1983).

Ignorable selection is a powerful assumption since it implies that conventional methods of adjustment can successfully remove bias due to nonrandom selection. Still, ignorable selection is also often a fairly tenuous assumption: although it is often not implausible, positive evidence in its support is usually lacking. It is, therefore, advisable to study the sensitivity of conclusions to plausible violations of ignorable selection (see Rosenbaum and Rubin 1983b; Rosenbaum 1984b) and to test the assumption whenever feasible (see Rosenbaum 1984c).

2. MODEL-BASED DIRECT ADJUSTMENT

2.1 Outline: An Estimator, Its Properties, and the Behavior of Estimated Propensity Scores

In this article, the directly adjusted estimator **D** and the sample mean **M** are viewed as two members of a class of estimators of $\bar{\mathbf{r}}$ based on different models for the selection probabilities or propensity scores. In effect, **D** involves estimating e_s by the sample proportion n_s/N_s, whereas **M** involves estimating e_s by a constant n_+/N for all s. Intuition suggests that it might be advantageous to allow estimates of selection probabilities to vary from subclass to subclass, so that the resulting estimator compensates for (ignorable) departures from simple random sampling, and yet to estimate e_s by a quantity that is more stable than the empirical proportion, n_s/N_s, particularly when the number of subclasses is large relative to N, and certainly when some $n_s = 0$. A short outline of the main points in the subsequent discussion follows.

1. In Section 2.2, model-based direct adjustment is defined and some of its attractive properties are developed. This discussion continues in the Appendix, where the variance of the estimator is examined.

2. Rubin and I have often observed that, in matching on propensity scores or subclassifying on propensity scores, *estimates* of the propensity score work better—that is, produce greater balance—than theory suggests *true* propensity scores should (Rosenbaum and Rubin 1983a, sec. 3.3, table 2; 1984, sec. 2.2; 1985, sec. 3). This sounds paradoxical at first, because in most contexts experience suggests that estimates of a parameter perform less well than true parameter values. This "paradox" is at least partially resolved in Section 2.3.

3. The practical performance of model-based direct adjustment is examined in Section 3.

2.2 Model-Based Weights for Subclass-Specific Means: The Estimator and Its Properties

Write ϕ_s for the logit of the selection probability or propensity score e_s, that is, $\phi_s = \log[e_s/(1 - e_s)]$, and write

$\boldsymbol{\phi}$ for $(\phi_1, \phi_2, \ldots, \phi_S)^T$. As has been done previously (Rosenbaum and Rubin 1984, 1985; Rosenbaum 1984a), the selection probabilities will be modeled using a logit model (Cox 1970) of the form

$$\boldsymbol{\phi} = \mathbf{F}\boldsymbol{\beta}, \qquad (2.1)$$

where $\mathbf{F}$ is a known $S \times P$ matrix of full column rank whose S rows describe features of the S subclasses and $\boldsymbol{\beta}$ is an unknown P-dimensional parameter, with $1 \le P \le S$. By a familiar argument (Cox 1970, sec. 2.3), $\mathbf{F}^T\mathbf{n}$ is sufficient for $\boldsymbol{\beta}$ in (2.1). An estimator of $\bar{\bar{\mathbf{r}}}$ that applies model-based weights to the subclass means $\bar{\mathbf{r}}_s$ is

$$\mathbf{L}_\mathbf{F} = \frac{1}{N}\sum_{s=1}^{S}\sum_{s=1}^{S}\frac{z_{si}\mathbf{r}_{si}}{\hat{e}_s} = \sum_{s=1}^{S}\frac{N_s}{N}\frac{n_s}{\hat{m}_s}\bar{r}_s, \qquad (2.2)$$

where $\hat{e}_s = \hat{m}_s/N_s$ is the maximum likelihood estimator (Cox 1970) of e_s under model (2.1) and $n_s \cdot \bar{\mathbf{r}}_s/\hat{m}_s$ is defined to be zero when $\hat{m}_s > 0$ but $n_s = 0$. Model-based direct adjustment has the following attractive properties.

(i) *A special case: the poststratified or directly adjusted estimator*. If $n_s > 0$ for all s, and if a fully saturated model is used in (2.1)—that is, if $\mathbf{F}$ is the $S \times S$ identity matrix $\mathbf{I}$—then $\hat{m}_s = n_s$ and $\hat{e}_s = n_s/N_s$, so $\mathbf{L}_\mathbf{I} = \mathbf{D}$. As noted previously, $\mathbf{D}$ is conditionally unbiased for $\bar{\bar{\mathbf{r}}}$ given $\mathbf{n}$ under ignorable selection—that is, $\mathbf{L}_\mathbf{I}$ is conditionally unbiased given the sufficient statistic $\mathbf{F}^T\mathbf{n} = \mathbf{I}^T\mathbf{n} = \mathbf{n}$ for $\boldsymbol{\beta}$.

(ii) *A special case: the sample mean*. If the selection probabilities are assumed constant—that is, if $\mathbf{F} = \mathbf{1}$ in (2.1) where $\mathbf{1}$ is an S-dimensional vector of 1's—then $\hat{m}_s = n_+N_s/N$ and $\hat{e}_s = n_+/N$, so $\mathbf{L}_\mathbf{1} = \mathbf{M}$. Again, when selection is ignorable and model (2.1) holds with $\mathbf{F} = \mathbf{1}$, the sample mean, $\mathbf{L}_\mathbf{1}$, conditionally unbiased for $\bar{\bar{\mathbf{r}}}$ given the sufficient statistic $\mathbf{F}^T\mathbf{n} = \mathbf{1}^T\mathbf{n} = n_+$ for $\boldsymbol{\beta}$.

(iii) *Superpopulation consistency*. If the N individuals in the finite population are themselves a simple random sample from an infinite superpopulation in which $\mathbf{r}$ has finite mean and variance in every subclass, and if selection is ignorable and model (2.1) holds, then $\mathbf{L}_\mathbf{F}$ is easily seen to be consistent for $\bar{\bar{\mathbf{r}}}$ in the superpopulation sense that $\mathbf{L}_\mathbf{F}$ and $\bar{\bar{\mathbf{r}}}$ both converge in probability to the superpopulation mean as $N \to \infty$. [Note that as $N \to \infty$ with model (2.1) fixed, $n_+ \to \infty$ a.s.]

(iv) *Model-based direct adjustment retains the two attractions of direct adjustment noted in Section 1.4 without the principal disadvantage*. Specifically, use of $\mathbf{L}_\mathbf{F}$ does not require modeling of the response $\mathbf{r}_{si}$ and produces parallel adjustments in the different coordinates of $\mathbf{r}_{si}$. Moreover, $\mathbf{D}$ is undefined if any n_s is zero, but $\mathbf{L}_\mathbf{F}$ is defined providing only that each fitted count $\hat{m}_s$ is greater than zero, a much weaker condition.

(v) *Larger weights for subclass means based on larger samples*. It is useful to contrast expressions (1.3) for $\mathbf{D}$ and (2.2) for $\mathbf{L}_\mathbf{F}$. In $\mathbf{D}$, the weight applied to $\bar{\mathbf{r}}_s$ is (N_S/N), whereas in $\mathbf{L}_\mathbf{F}$ the weight is $(N_S/N)(n_s/\hat{m}_s)$. Note first that if the model (2.1) provides an adequate fit, perhaps judged by a chi-squared statistic, then $n_s/\hat{m}_s = (n_s/N_s)/\hat{e}_s$ will typically be close to 1, particularly for subclasses in which N_s is large and e_s is not too small—otherwise, if this were untrue for all but a few subclasses, the fit would be poor. Still, compared with $\mathbf{D}$, $\mathbf{L}_\mathbf{F}$ gives slightly greater weight to subclasses in which $n_s > \hat{m}_s$, and slightly lower weight to subclasses with $n_s < \hat{m}_s$, that is, greater weight to subclasses in which the sample mean $\bar{\mathbf{r}}_s$ is based on a larger sample size than model (2.1) predicted. In the extreme, when $n_s = 0$ so there are no observations available to estimate $\bar{\mathbf{r}}_s$, the model-based estimator assigns zero weight to the subclass.

(vi) *A nearly equivalent estimator that is conditionally unbiased*. Since, assuming (1.4), n_s is unbiased for $m_s = E(n_s) = N_se_s$, and since $\mathbf{F}^T\mathbf{n}$ is a complete, sufficient statistic for $\boldsymbol{\beta}$ under model (2.1), it follows from the Rao–Blackwell theorem (e.g., Rao 1973, p. 326) that $\tilde{\mathbf{m}} = E(\mathbf{n} \mid \mathbf{F}^T\mathbf{n})$ is a minimum variance unbiased estimator under (2.1) of $\mathbf{m} = (m_1, m_2, \ldots, m_S)^T$. The minimum variance estimator $\tilde{\mathbf{m}}$ is difficult to compute except in special cases, but it possesses certain interesting theoretical properties and behaves very much like the maximum likelihood estimator $\hat{\mathbf{m}}$. In particular, in the two special cases (i) and (ii), $\tilde{\mathbf{m}} = \hat{\mathbf{m}}$. If $\tilde{\mathbf{m}}$ is used in place of the maximum likelihood estimator $\hat{\mathbf{m}}$ in (2.2), then the resulting conditional estimator, $\mathbf{C}_\mathbf{F}$ say, is conditionally unbiased for $\bar{\bar{\mathbf{r}}}$ given the sufficient statistic $\mathbf{F}^T\mathbf{n}$ for $\boldsymbol{\beta}$, providing of course that selection is ignorable and model (2.1) holds. To see this, evaluate $E\{\mathbf{C}_\mathbf{F} \mid \mathbf{F}^T\mathbf{n}\}$ in two stages as

$$E\{E(\mathbf{C}_\mathbf{F} \mid \mathbf{n}) \mid \mathbf{F}^T\mathbf{n}\} = E\left\{\sum_{s=1}^{S}\frac{N_s}{N}\frac{n_s}{\tilde{m}_s}\bar{\mathbf{r}}_s \;\middle|\; \mathbf{F}^T\mathbf{n}\right\} = \bar{\bar{\mathbf{r}}}.$$

As discussed in Section 1.4 and in greater detail by Holt and Smith (1979), conditional unbiasedness is often a property of greater practical relevance than unbiasedness or consistency, since an estimator that is conditionally unbiased is known to behave sensibly in samples that resemble the sample actually obtained. It is a consequence of Haberman's (1974, p. 76) theorem 4.1 in the conditional Poisson case that $(1/N)(\hat{\mathbf{m}} - \tilde{\mathbf{m}}) = o_p(1/\sqrt{N})$ as $N \to \infty$, so the use of $\hat{\mathbf{m}}$ in place of $\tilde{\mathbf{m}}$ will often have a negligible effect in large samples from large populations. [Specifically, this follows from the familiar relationship between log-linear and logit models and from Haberman's expression (4.9) with his subspace M equal to his subspace N, in which case his limiting normal distribution is degenerate.] More precisely, the conditional bias of $\mathbf{L}_\mathbf{F}$ is of a smaller stochastic order as $N \to \infty$ than the conditional variance of $\mathbf{L}_\mathbf{F}$ (see the Appendix)—that is, informally the asymptotic conditional bias is negligible.

2.3 A Digression: Estimated Versus True Propensity Scores

This section relates the current article to previous work and may be skipped without loss of continuity. As noted in Section 2.1, experience has shown that matching on *estimated* propensity scores or subclassification on *estimated* propensity scores tends to produce greater control of imbalances in the variables used to construct the propensity score than theory suggests *true* propensity scores should produce (Rosenbaum and Rubin 1983a, secs. 2.3 and 3.3; 1984, sec. 2.2; 1985, sec. 3). From a practical perspective, matching and subclassification on estimated propensity scores are often attractive methods of adjusting for the propensity score: they are easy to implement, easy to interpret, often convincing to nontechnical audiences, and they can easily accommodate additional model-based adjustments (e.g., Rosenbaum and Rubin 1984, sec. 3.3). Still, matching and subclassification involve rather coarse, algorithmic adjustments for the propensity score, which tend to obscure the reason why estimated propensity scores outperform true propensity scores. In contrast, this issue is clearer in terms of model-based direct adjustment. Write **T** for the estimator obtained by substituting the true propensity score e_s and expected count $m_s = N_s e_s$ for the estimated propensity score $\hat{e}_s$ and fitted count $\hat{m}_s$ in (2.2).

Consider, first, the simplest case, namely selecting individuals by tossing a coin N times, with Pr(head) $= e^*$, and selecting individual i if a head is obtained on the ith toss. The estimator, **T**, based on the true propensity score is

$$\frac{1}{Ne^*} \sum_{s=1}^{S} \sum_{i=1}^{N_s} z_{si} \mathbf{r}_{si},$$

that is, the sample total divided by the expected sample size, Ne^*. Clearly, **T** is unbiased, but it is conditionally biased given the total sample size n_+. For example, if the coordinates of **r** were always nonnegative, then **T** would have a positive conditional bias in each coordinate when $n_+ > Ne^*$ and a negative conditional bias when $n_+ < Ne^*$; that is, we could estimate the sign of the error of **T**, namely $\mathbf{T} - \bar{\mathbf{r}}$, using n_+. In contrast, had we estimated $e^* = \Pr(z_{si} = 1)$ under the logit model that fits the same probability of selection for all subclasses—that is, the model with $\mathbf{F} = \mathbf{1}$—the resulting estimator $\mathbf{L}_1$ is the sample mean **M**, which is conditionally unbiased given n_+. In this simplest case, the estimator based on an estimated propensity score is clearly preferable to an estimator based on the true propensity score. Indeed, overfitting the propensity score by using a saturated model in (2.1)—that is, a model with $\mathbf{F} = \mathbf{I}$—yields direct adjustment, $\mathbf{L}_\mathbf{I} = \mathbf{D}$, which is often superior to the sample mean. In other words, overfitting of the propensity score is not necessarily harmful.

In general **T** is the Horvitz–Thompson (1952) estimator, which is unbiased for $\bar{\mathbf{r}}$ but not conditionally unbiased given functions of **n**. In constrast, the estimator, $\mathbf{C}_\mathbf{F}$, based on an unbiased estimator of the propensity score, is conditionally unbiased given $\mathbf{F}^T\mathbf{n}$, and the estimator $\mathbf{L}_\mathbf{F}$ based on maximum likelihood is nearly equivalent to $\mathbf{C}_\mathbf{F}$ [see Sec. 2.2(vi)]. Write $m_+ = m_1 + m_2 + \cdots + m_S$. Informally, the superiority of estimators involving estimated propensity scores in place of the true propensity scores results because **T** compensates only for the systematic difference between the population and expected sample proportions—N_S/N and m_s/m_+—whereas $\mathbf{C}_\mathbf{F}$ and $\mathbf{L}_\mathbf{F}$ compensate to some degree for the difference between the population and actual sample proportions—N_s/N and n_s/n_+—thereby correcting for both systematic and chance imbalances.

3. USING MODEL-BASED DIRECT ADJUSTMENT TO ESTIMATE A DISTRIBUTION FUNCTION FROM A SELECTED SAMPLE

3.1 Estimating the Population Distribution of Essay 5 Scores From the Subgroup of Examinees Who Chose To Write Essay 5

As discussed in Section 1.2, model-based direct adjustment will be used in an effort to estimate the distribution of Essay 5 scores that would have been obtained if all students taking the Advanced Placement Biology Exam had been required to write Essay 5. In principle, if selection were ignorable, either conventional or model-based direct adjustment could be used; in fact, however, only model-based adjustment can be applied in the current example, because there are many sparse and many empty subclasses.

Subclasses were constructed from functions of the item responses that appeared predictive of essay choice, as judged from several stepwise regressions. Specifically, the subclasses were based on (Subscore 1 + Subscore 3), the number of omitted items, and the binary correct/incorrect responses to items 4, 8, 37, 52, 61, and 51. Based on Cochran's (1968) results, the Subscore 1 + Subscore 3 total and the number of omitted items were each grouped into four categories, resulting in $4^2 \times 2^6 = 1{,}024$ potential subclasses and 915 actual subclasses containing at least one examinee. This process of subclass construction is essentially subclassification on a crude balancing score (e.g., Rosenbaum and Rubin 1983a; 1984). Of these 915 subclasses, 165 contained no examinee with a score on Essay 5, so conventional direct adjustment could not be applied.

3.2 Estimating the e_s's

The propensity scores or subclass-specific probabilities of selecting Essay 5—that is, the e_s's in (1.4) and (2.1)—were estimated by maximum likelihood under a logit model that included six binary variables for Items 4, 8, 52, 37, 61, and 51, and three indicator variables each for the four categories of (Subscore 1 + Subscore 3) and the number of omitted items. Two interactions, between Items 4 and 51 and between Items 8 and 61, were added on the basis of a stepwise selection of interactions.

Since the $\hat{e}_s$'s are transformed to serve as weights for individual observations in (2.2), it is natural to think of $\hat{e}_s$ as attached to each examinee in subclass s—that is, $i = 1, 2, \ldots, N_s$—and to examine the distribution of the 15,676 $\hat{e}_s$'s for the 15,676 examinees. Table 1 contains five-number summaries (Tukey 1977) of the $\hat{e}_s$'s within eight large groups of examinees formed from two important covariates, Subscores 1 + 3 and Item 4. The estimated probabilities of selecting Essay 5 range from a low of .11 to a high of .55. In contrast, the estimates of the $\hat{e}_s$'s under the saturated model [i.e., $\mathbf{F} = \mathbf{I}$ in (2.1)] that is implicit in the use of conventional direct adjustment would have ranged from 0 to 1. Using $\mathbf{L}_\mathbf{F}$, an examinee who selected Essay 5 would receive a weight of $1/\hat{e}_s$ in estimating the population distribution of Essay 5 scores, so these weights

Table 1. Five-Number Summaries of $\hat{e}_i$'s for 15,676 Examinees, by Quartile on Subscores 1 + 3 and by Item 4

Quartile on subscores 1 + 3	Item 4 Incorrect/Omit			Item 4 Correct		
Lowest						
M		.23			.14	
Q	.20		.25	.13		.16
E	.16		.36	.11		.23
Second						
M		.30			.20	
Q	.27		.34	.18		.22
E	.21		.44	.14		.29
Third						
M		.35			.23	
Q	.32		.38	.21		.25
E	.24		.48	.16		.33
Highest						
M		.47			.32	
Q	.45		.51	.31		.36
E	.31		.55	.22		.41

NOTE: M = median; Q = quartiles; E = extremes.

are contained in the range from 1/.55 = 1.8 to 1/.11 = 9.1. In contrast, the saturated model associated with conventional direct adjustment would yield weights ranging from 1 to ∞, with the infinite weights attached to subclasses in which all examinees selected Essay 6. In short, the weights associated with $\mathbf{L}_\mathbf{F}$ under the current model do vary—as weights should—from examinee to examinee, but they vary much less than the weights that would be implicit in the use of conventional direct adjustment.

3.3 Evaluating $\mathbf{L}_\mathbf{F}$

To evaluate the performance of $\mathbf{L}_\mathbf{F}$, the population proportion of examinees answering each multiple-choice item correctly will be estimated twice by using $\mathbf{L}_\mathbf{F}$, once from examinees selecting Essay 5 and once from those selecting Essay 6: the difference of these two estimates for each item should estimate zero. The boxplots in Figure 2 display the results. For comparison, the first boxplot displays the 120 differences in proportions prior to adjustment (the unstandardized $p_5 - p_6$'s in Fig. 1). The second boxplot displays the 120 adjusted differences based on applying $\mathbf{L}_\mathbf{F}$ twice and differencing, namely

$$\mathbf{d} = \frac{1}{N}\left\{\sum_{s=1}^{S}\sum_{i=1}^{N_s}\frac{z_{si}\mathbf{r}_{si}}{\hat{e}_s} - \sum_{s=1}^{S}\sum_{i=1}^{N_s}\frac{(1-z_{si})\mathbf{r}_{si}}{(1-\hat{e}_s)}\right\}, \quad (3.1)$$

where $\mathbf{r}_{si}$ is in this context the 120-dimensional vector of binary item responses. Adjustment by (2.2) or, equivalently, (3.1) reduced the median difference (i.e., median $p_5 - p_6$) from .045 to .005. The spread of the differences also decreased; in particular, the extreme negative difference for Item 4 was reduced from −.069 to .0007. As one might expect, *the results for the six items specifically included in the subclassification were generally superior to the results for many other items* in Figure 2: in particular, the differences for Items 4, 8, 37, 52, 61, and 51 were −.0687, .0045, .0990, .0908, .0762, and .0891 prior to adjustment, and .0007, −.0021, .0006, .0027, −.0008, and −.0011 after adjustment. The sum of the absolute values of these six differences was .428 prior to adjustment and .008 after adjustment, yielding a reduction of 100(1 − .008/.428) = 98%.

3.4 Applying $\mathbf{L}_\mathbf{F}$

Essay 5 was scored on a 15-point raw scale. The empirical cumulative distribution of essay scores for examinees who actually wrote Essay 5 is given in Table 2 in the

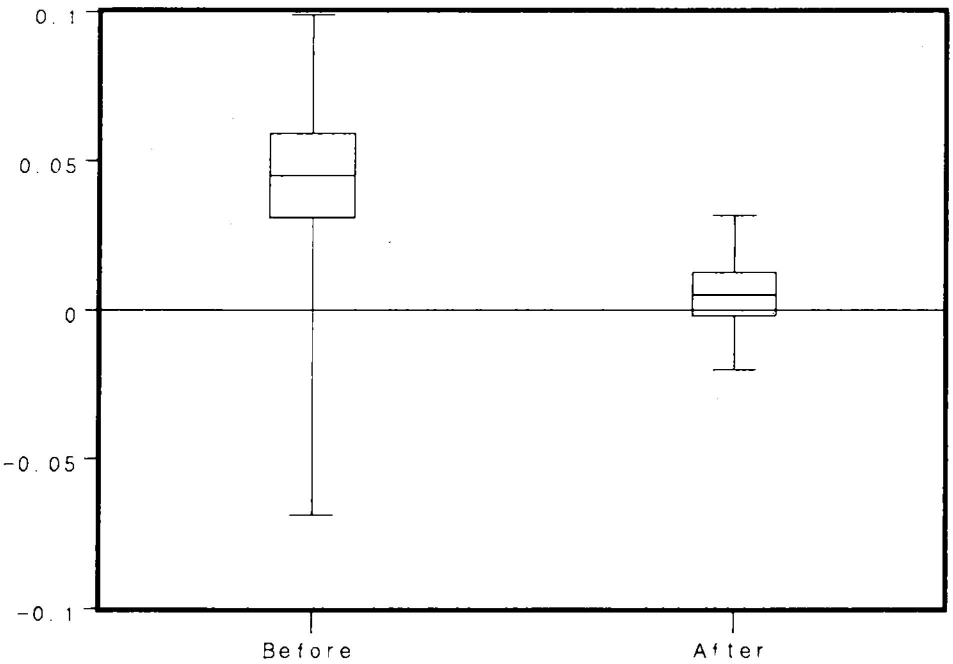

Figure 2. Boxplots of 120 Differences in Proportions of Examinees Answering Items Correctly Before and After Adjustment.

Table 2. Cumulative Distributions of Essay Scores Before (B) and After (A) Adjustment

Essay score	Essay 5	
1	B	.14
	A	.16
2	B	.30
	A	.34
3	B	.47
	A	.51
4	B	.63
	A	.67
5	B	.74
	A	.78
6	B	.83
	A	.86
7	B	.90
	A	.92
8	B	.94
	A	.95
9	B	.96
	A	.97
10	B	.981
	A	.984
11	B	.992
	A	.993
12	B	.996
	A	.997
13	B	.999
	A	.999
14	B	1.000
	A	.999
15	B	1.000
	A	1.000

lines labeled "B" for before adjustment. Also in Table 2 is the adjusted estimate of the distribution of Essay 5 scores for all examinees, adjusted by applying $\mathbf{L_F}$ to the 15-dimensional binary vector, $\mathbf{r}_{si}$ say, whose jth coordinate indicates whether the ith student subclass s has a score less than or equal to j. The adjusted distribution of Essay 5 scores has, for the most part, been shifted to the left of the unadjusted distribution, suggesting that, had all students written Essay 5, the distribution of scores would have been lower than the observed Essay 5 score distribution.

APPENDIX: THE VARIANCE OF $\mathbf{L_F}$

The most natural variance to associate with $\mathbf{L_F}$ is the conditional variance given the sufficient statistic $\mathbf{F}^T\mathbf{n}$ for the nuisance parameter $\boldsymbol{\beta}$. In the special cases of the sample mean (i.e., $\mathbf{F} = \mathbf{1}$) and the directly adjusted estimator (i.e., $\mathbf{F} = \mathbf{I}$), this definition of the variance of $\mathbf{L_F}$ recovers the usual formulas with finite population corrections [e.g., for the sample mean, the formula given by Cochran (1977), th. 2.2, p. 23; for the directly adjusted estimator, the formula given by Holt and Smith (1979), expression 2]. Note that the maximum likelihood estimator $\hat{\mathbf{m}}$ is a function of the sufficient statistic $\mathbf{F}^T\mathbf{n}$ and is, therefore, fixed by the conditioning. Clearly,

$$\begin{aligned}
&\operatorname{var}(\mathbf{L_F} \mid \mathbf{F}^T\mathbf{n}) \\
&\quad = E\{\operatorname{var}(\mathbf{L_F} \mid \mathbf{n}) \mid \mathbf{F}^T\mathbf{n}\} + \operatorname{var}\{E(\mathbf{L_F} \mid \mathbf{n}) \mid \mathbf{F}^T\mathbf{n}\} \\
&\quad = E\left\{\sum_{s=1}^{S} \left(\frac{N_s n_s}{N\hat{m}_s}\right)^2 \frac{\mathbf{V}_s}{n_s}\left(1 - \frac{n_s}{N_s}\right) \,\middle|\, \mathbf{F}^T\mathbf{n}\right\} \\
&\qquad + \operatorname{var}\left\{\sum_{s=1}^{S} \frac{N_s n_s}{N\hat{m}_s}\bar{\bar{\mathbf{r}}}_s \,\middle|\, \mathbf{F}^T\mathbf{n}\right\} \\
&\quad = \sum_{s=1}^{S} \left(\frac{N_s}{N}\right)^2 \left(\frac{\bar{m}_s}{\hat{m}_s}\right) \frac{\mathbf{V}_s}{\hat{m}_s}\left(1 - \frac{h_s}{N_s}\right) \\
&\qquad + \sum_{s=1}^{S}\sum_{t=1}^{S} \frac{N_s N_t}{N^2 \hat{m}_s \hat{m}_t} \bar{\bar{\mathbf{r}}}_s \bar{\bar{\mathbf{r}}}_t^T \operatorname{cov}(n_s, n_t \mid \mathbf{F}^T\mathbf{n}), \qquad \text{(A.1)}
\end{aligned}$$

where

$$\mathbf{V}_s = (N_s - 1)^{-1} \sum_{i=1}^{N_s} (\mathbf{r}_{si} - \bar{\bar{\mathbf{r}}}_s)(\mathbf{r}_{si} - \bar{\bar{\mathbf{r}}}_s)^T$$

and

$$h_s = E(n_s^2 \mid \mathbf{F}^T\mathbf{n})/\bar{m}_s.$$

Both sums in (A.1) are nonnegative definite matrices: they are, in a sense, the within- and between-subclass components of the variance. The second sum is zero for the directly adjusted or poststratified estimator, $\mathbf{L_I} = \mathbf{D}$, provided that each n_s is greater than zero; for this specific estimator, however, there is a division by zero in both sums—in effect, an infinite variance—when any $n_s = 0$, since in this case $\hat{m}_s = n_s$. The finite population corrections in the first term of (A.1) are on an average larger numbers—that is, smaller downward corrections—for the directly adjusted estimator for any other estimator of the form (2.2), since for any $\mathbf{F}$

$$(1 - h_s/N_s) \le (1 - \bar{m}_s/N_s) = E(1 - n_s/N_s \mid \mathbf{F}^T\mathbf{n}),$$

where $(1 - n_s/N_s)$ is the correction for subclass s when $\mathbf{L_I} = \mathbf{D}$ is used.

The covariance matrix, $\operatorname{cov}(\mathbf{n} \mid \mathbf{F}^T\mathbf{n})$, involved in the second term of (A.1) behaves in a predictable manner as the model (2.1) is changed. Suppose that $\mathbf{F} = (\mathbf{A}, \mathbf{B})$ and we wish to compare $\operatorname{cov}(\mathbf{n} \mid \mathbf{F}^T\mathbf{n})$ for $\mathbf{L_F}$ with $\operatorname{cov}(\mathbf{n} \mid \mathbf{A}^T\mathbf{n})$ for $\mathbf{L_A}$. Clearly,

$$\operatorname{cov}(\mathbf{n} \mid \mathbf{A}^T\mathbf{n}) = E\{\operatorname{cov}(\mathbf{n} \mid \mathbf{F}^T\mathbf{n})|\mathbf{A}^T\mathbf{n}\} + \operatorname{cov}\{E(\mathbf{n} \mid \mathbf{F}^T\mathbf{n})|\mathbf{A}^T\mathbf{n}\},$$

so

$$\operatorname{cov}(\mathbf{n} \mid \mathbf{A}^T\mathbf{n}) - \operatorname{cov}(\mathbf{n} \mid \mathbf{F}^T\mathbf{n})$$

is a matrix that has a nonnegative definite conditional expectation given $\mathbf{A}^T\mathbf{n}$—that is, more elaborate models tend to make the conditional covariance matrix of $\mathbf{n}$ smaller. A general large sample approximate for $\operatorname{cov}(\mathbf{n} \mid \mathbf{F}^T\mathbf{n})$ may be obtained from Haberman's (1974, p. 18) theorem 1.1.

There are, then, several forces working against one another in (A.1) as the model is varied. In addition, (A.1) depends on the way the population mean responses $\bar{\bar{\mathbf{r}}}_s$ vary from subclass to subclass. Finally, the interpretation of the comparisons of conditional variances of $\mathbf{L_F}$ given $\mathbf{F}^T\mathbf{n}$ for different $\mathbf{F}$'s are, to an extent, complicated by the fact that both the estimator and the conditioning event are changing. In the case of simple random sampling, Holt and Smith (1979) addressed this last point in their comparison of $\mathbf{M}$ and $\mathbf{D}$ by conditioning throughout on $\mathbf{n}$—that is, they compared the conditional variance given $\mathbf{n}$ of $\mathbf{D}$ and the conditional mean squared error given **n** of **M**— finding that neither estimator strictly dominates the other.

[*Received July 1984. Revised May 1986.*]

REFERENCES

Cochran, W. G. (1968), "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24, 295–313.

——— (1977), *Sampling Techniques*, New York: John Wiley.

Coffman, W. E. (1971), "Essay Examinations," in *Educational Measurement*, ed. R. L. Thorndike, Washington, DC: American Council on Education, pp. 271–302.

Cox, D. R. (1970), *The Analysis of Binary Data*, London: Methuen.

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

——— (1979), "Conditional Independence in Statistical Theory" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 41, 1–32.

Fleiss, J. L. (1981), *Statistical Methods for Rates and Proportions* (2nd ed.), New York: John Wiley.

Haberman, S. (1974), *The Analysis of Frequency Data*, Chicago: Chicago University Press.

Holt, D., and Smith, T. M. F. (1979), "Post Stratification," *Journal of the Royal Statistical Society*, Ser. A, 142, 33–46.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Little, R. J. A. (1982), "Models for Nonresponse in Sample Surveys," *Journal of the American Statistical Association*, 77, 237–250.

——— (1984), "Survey Nonresponse Adjustments," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 1–10.

Mosteller, F., and Tukey, J. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley.

Oh, H. L., and Scheuren, F. J. (1983), "Weighting Adjustment for Unit Nonresponse," in *Incomplete Data in Sample Surveys* (Vol. 2), eds. W. Madow, I. Olkin, and D. B. Rubin, New York: Academic Press, pp. 143–184.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley.

Rosenbaum, P. R. (1984a), "Conditional Permutation Tests and the Propensity Score in Observational Studies," *Journal of the American Statistical Association*, 79, 565–574.

——— (1984b), "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment," *Journal of the Royal Statistical Society*, Ser. A, 147, Part 5, 656–666.

——— (1984c), "From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment," *Journal of the American Statistical Association*, 79, 41–48.

——— (1984d), "Testing the Conditional Independence and Monotonicity Assumptions of Item Response Theory," *Psychometrika*, 49, 425–436.

——— (1985), "Comparing Distributions of Item Responses for Two Groups," *British Journal of Mathematical and Statistical Psychology*, 38, 206–215.

Rosenbaum, P. R., and Rubin, D. B. (1983a), "The Central Role of the Propensity Score in Observational Studies for Casual Effects," *Biometrika*, 70, 41–55.

——— (1983b), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45, 212–218.

——— (1984), "Reducing Bias in Observational Studies Using Subclassification of the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.

——— (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 39, 33–38.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

——— (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1–26.

——— (1978), "Bayesian Inference for Casual Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.

——— (1983), Comment on "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," by M. H. Hansen, W. G. Madow, and B. J. Tepping, *Journal of the American Statistical Association*, 78, 803–805.

——— (1984), Comment on "Graphical Methods for Assessing Logistic Regression Models," by J. M. Landwehr, D. Pregibon, and A. C. Shoemaker, *Journal of the American Statistical Association*, 79, 79–80.

Smith, T. M. F. (1983), "On the Validity of Inference From Non-random Samples," *Journal of the Royal Statistical Society*, Ser. A, 146, 394–403.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.