

ANALYTICAL MODELING IN COMPLEX SURVEYS OF WORK PRACTICES

JEROME P. REITER, ELAINE L. ZANUTTO, and LARRY W. HUNTER*

*Jerome P. Reiter is Assistant Professor of the Practice of Statistics and Decision Sciences at Duke University. Elaine L. Zanutto is Assistant Professor of Statistics in the Wharton School at the University of Pennsylvania. Larry W. Hunter is Assistant Professor of Management and Human Resources in the School of Business at the University of Wisconsin-Madison. They are grateful to Paul Osterman for generously sharing data from the 1992 National Survey of Establishments. They also appreciate the help of Arne Kalleberg and Jerry Jacobs in acquiring and using the data from the National Organizations Survey. They thank Jed Frees for his comments. The data, and the statistical programs (from Stata 7.0) used to generate the results in this study, can be obtained from Larry Hunter, assuming approval of the original supplier of the data.

Abstract

Quantitative industrial relations research frequently relies upon data collected from large surveys of establishments that use complex sampling designs, such as stratified and unequal probability sampling. In this paper, we analyze two complex surveys of establishments, the National Organizations Survey and the National Survey of Establishments. We discuss model-based and design-based, i.e. survey-weighted, strategies for analyzing these data. We show that the choice of strategy can affect inferences about parameters, and hence conclusions drawn from analyses. We discuss the advantages of model-based approaches that include independent variables that correspond to design features, such as functions of size measures or indicator variables for strata or clusters, relative to purely design-based approaches.

Much contemporary research in industrial relations draws on large surveys with complex sampling designs. In the U.S., for example, data from sources such as the Current Population Survey and National Longitudinal Survey of Youth have long been used to draw inferences and test hypotheses about populations of individual workers. In the past decade, survey research has been extended to include large samples in which establishments or organizations serve as the units of analysis. Both single-industry surveys and more broadly representative, cross-industry samples have provided data for a number of studies of the determinants of variation in human resource practices as well as of the effects of practices on key performance outcomes.

Many of the early surveys of organizations drew on samples of convenience, making few or no claims to statistical representativeness. But more recent studies have used such techniques as stratification and unequal probability sampling to increase accuracy of population-level estimates while keeping costs manageable. The data sets associated with these complex surveys often provide detailed descriptions of the design, as well as further information about the design in the form of survey weights attached to individual observations.

It is generally accepted that estimation of descriptive statistics, such as population means and totals, in complex surveys should account for the design, as is typically done via survey-weighted estimators. Of course, descriptive statistics are only one aspect of research on industrial relations. Researchers often seek inferences for analytical estimands, such as the coefficients in regression models used to consider the determinants and effects of work practices. For these estimands, the role of the survey design is less clear, resulting in approaches to analytical modeling that range from no explicit attention to the design to strict adherence to design-based (i.e. survey-weighted) analyses.

Should information about complex sampling designs affect the formulation and estimation of analytical models used to consider the determinants and effects of work practices? If so, how should the design information be incorporated into analyses? And, how does the approach used to account for complex survey designs affect the substantive conclusions drawn

from these analyses?

Our paper addresses these questions by presenting analyses of two different complex establishment-level surveys. Each of the two analyses provides a specific example of a general question that has attracted the interest of industrial relations researchers: what differentiates firms using particular work practices from those that do not use those practices? Using the National Organizations Survey, we investigate the correlates of firm-level internal labor market structures. Using the National Survey of Establishments, we analyze factors associated with the use of self-directed work teams.

In our analyses, we illustrate and compare two approaches to incorporating the sampling design in analytical models, namely (i) a design-based approach that uses survey weights to estimate model parameters and (ii) a model-based approach that includes design features as independent variables in models, using standard likelihood-based methods to estimate model parameters. The analyses illustrate that substantive conclusions about the correlates of firm internal labor markets, and work teams, can be affected by the choice of approach. In both analyses, we found that the model-based approach enabled us to find interactions and non-linear relationships that we may have overlooked when using the design-based approach, indicating an advantage of the model-based approach. By describing our model-building process in detail, and how that process is driven in part by the sampling design, we ultimately provide a template that is generalizable beyond the two examples we have chosen. The template is intended to help researchers obtain better-fitting models, and hence reach more valid conclusions, when analyzing complex surveys.

Firm-level and Establishment-level Survey Design and Analysis

The use of firm-level and establishment-level surveys in industrial relations and human resources research became more common in the 1990s, especially because they permitted

large-sample analyses of questions associated with the connections between work practices, business strategy, and economic performance (Hunter and Pil, 1995). Some early influential papers drew on samples of convenience, often limited to a single industry. MacDuffie (1995), for example, compared work practices and productivity across 62 auto assembly plants that represented about half the capacity of the industry. Other studies sought comprehensive coverage of a broader range of firms: Huselid (1995), in analyzing connections between human resource practices and firm performance, mailed questionnaires to all 3,452 domestically owned, independent firms that had filed 10-K reports in 1991-1992.

These kinds of approaches, however, had drawbacks. Results from industry studies could not be readily generalized. Sampling only firms with publicly available financial data missed many employers, especially small ones. Further, measuring practices at the firm-level rather than at the establishment-level may have created additional measurement error (Gerhart *et al.*, 2000). Thus, researchers began using complex sampling designs, such as unequal probability sampling and stratified sampling (see Cochran, 1977). The National Employers Survey administered by the U.S. Bureau of the Census (Cappelli, 2001), for example, sampled private, for-profit work establishments of 20 workers or more, oversampling large establishments and manufacturers. Gramm and Schnell (2001), to take a second, contrasting example of a complex survey, did not seek a nationally representative sample. Rather, they surveyed Alabama employers in manufacturing and scrap processing, oversampling those with high-technology products.

Estimating descriptive statistics for the populations in these kinds of surveys requires accounting for the design. Such accounting is typically accomplished by using estimators that employ survey weights determined by the sample design, such as Horvitz-Thompson estimators (Horvitz and Thompson, 1952). Use of these weights is required in order to obtain unbiased estimates of population and group means and associated standard errors (Cochran, 1977). Estimators that do not account for the design, such as unweighted means, may yield badly biased estimates if the intent of such estimates is to generalize to the population

of establishments. Kalleberg *et al.* (1996b), for example, demonstrated large differences in the National Organizations Study between weighted and unweighted estimates of American establishment characteristics and usage of work practices.

The estimation of descriptive population statistics is typically only one of the purposes toward which researchers direct analyses of surveys of work practices. Such surveys also contain data that are suitable for multivariate analyses. For example, studies have drawn on establishment surveys to identify factors associated with the adoption of firm-level internal labor market structures (e.g., Pfeffer and Cohen, 1984) and the adoption of innovative work practices (e.g., Osterman, 1994, 2000; Pil and MacDuffie, 1996).

In contrast to the clarity associated with the rationale for using weights to produce descriptive statistics, there is no consensus about the role of survey weights, or the sampling design, in multivariate analyses of survey data of work practices. Gramm and Schnell (2001), for example, use survey-weighted estimates derived from pseudo-likelihood methods, offering an explicit, design-based justification for weight-based estimation. Lynch and Black (1998), Cappelli and Neumark (2001), and other work featuring multivariate analyses of the Census Bureau's NES, in contrast, relied upon estimation of unweighted regression models. This lack of consensus motivates our consideration of how best to account for the survey design in analytical modeling.¹

Factors Associated with the Adoption of Work Practices

We illustrate the analysis of complex surveys using two well-known data sets, the National Organizations Study (NOS) and the National Survey of Establishments (NSE). Each survey has yielded data that have provided the basis for published analyses of work practices and their determinants. Each survey seeks representativeness through its complex design: the

¹This issue differs from the issues addressed by selection models (see Heckman, 1979, 1980). Selection models concern the biases caused by unknown, nonrandom selection mechanisms that are related to the outcome of interest. This requires estimating selection probabilities from the data. In the survey sampling context, sampling probabilities are known and fixed by the sample design.

NOS (Spaeth and O’Rourke, 1996) selected work establishments with probability proportional to their size, as measured by the number of employees; the NSE (Osterman, 1994) stratified across size and industry categories to create adequate samples for each of the categories, limiting the sample to establishments employing 50 or more workers. In this section, we describe the NOS and NSE, and we outline, for each survey, a question that is amenable to multivariate analyses with the data.

The National Organizations Study and “Firm Internal Labor Markets”

The National Organizations Study was a multi-investigator research project designed to study various human resources practices and other organizational characteristics of U.S. work establishments. Establishments were sampled with probability proportional to their number of employees. Nearly a quarter of the establishments in the survey have ten or fewer employees. The largest establishments have many more: over 13% of the sample employ more than 1,000 workers, and five of the respondents employ more than 10,000. The weights, therefore, vary by four to five orders of magnitude. The NOS was unusual in collecting data from tiny establishments; most surveys deliberately omit them.²

The NOS data have been used to fit many multivariate analyses. Following Kalleberg *et al.* (1996a), we analyze the correlates of “Firm Internal Labor Market” (FILM) structures. Understanding the factors associated with FILMs has been a question of long-standing research interest dating back at least to Doeringer and Piore (1971); Althauser and Kalleberg (1981), Osterman (1984), and Althauser (1989) provide reviews. This is because FILMs in turn are often correlated with wage outcomes, training opportunities, fringe benefits, and other human resource practices that affect workers in the labor market.

Our analyses seek to identify factors associated with adoption of FILMs by use of re-

²See Spaeth and O’Rourke (1996) for a thorough description of the issues associated with the design of the survey.

gression modeling. Similar analyses were done by Kalleberg *et al.* (1996a), and we replicate and extend their analysis. The dependent variable for our analyses, which we label *FILMscore*, is a mean score for an establishment on various survey questions. High values of *FILMscore* indicate establishments with highly developed firm internal labor markets: systems of job classification, job ladders, and internal promotion opportunities. Low values of *FILMscore* indicate that firms tend to hire from outside, and offer few opportunities for advancement. Following previous analyses, our independent variables are factors thought to be associated with more extensive use of FILMs: the number of employees in the establishment (*size*); the natural log of the number of hierarchical levels in the establishment (*lnlev*); the number of different departments in the organization (*depts*); survey-derived measures of decentralization (*decent*) and formalization (*formal*); dummy variables for establishments that produce services only (*service*), and for establishments that produce both products and services (*prodserv*); a survey item indicating the geographic scope of the establishment's target market (*scope*); the natural log of the age of the establishment (*lnage*); a survey-derived scale indicating the extent of problems in attracting and retaining employees (*eeprob*); a scale indicating the complexity of the establishment's environment (*complex*); dummy variables for public (*public*) and non-profit (*nonprofit*) establishments; scales measuring institutionalization (*instn*) and pressure from trade unions (*union*); and a dummy variable for establishments that are members of multi-site organizations (*multisite*).

The analysis of the NOS by Kalleberg *et al.* (1996a) shows that FILMs are associated with nearly all of these factors: the bivariate correlations tend to be statistically significant. Multivariate analysis of the same data, however, suggest that only a few of these factors are associated with FILMs when controlling for the others: establishment size and the number of hierarchical levels; the extent of formalization; membership in a multi-site organization; and, the production of both services and products. We extend this analysis by considering explicitly the extent to which different ways of treating the complexity of the design of the NOS affect these multivariate estimates.

The sampled data comprise 688 observations, of which only 573 have complete data on the variables of interest. Rates of missing data for individual variables range from 0% to 6%. Using only the 573 complete cases in analyses throws out valuable information on the 115 cases with only partially missing data. To avoid wasting this information, we used multiple imputation for missing data (Rubin, 1987), which essentially fills in values of missing items with predictions from models fit using the observed data. We generated imputations from multivariate normal models using the publicly available software “norm” written by Schafer (1997). Imputations for integer-valued variables were rounded to the nearest integer (Schafer, 1997, p. 202). We independently imputed five separate plausible values for each missing datum to create five completed data sets. These completed data sets were then analyzed individually, and the results were combined using the software “miest” written by Gary King (<http://gking.harvard.edu>). Further discussion of multiple imputation is provided later.

The NSE and self-directed work teams

The National Survey of Establishments was designed to yield estimates of the prevalence of a wide range of human resource practices in American firms; to enable researchers to examine the factors associated with the adoption of these practices; and, through follow-up surveys, to examine diffusion of innovative practices. In contrast to the NOS, the NSE covered only private sector for-profit businesses that employed at least fifty workers. Businesses were selected randomly from eighteen strata formed by a cross-classification of establishments into nine size categories based on numbers of employees, and two industry categories, manufacturing and non-manufacturing sectors. Sampling rates were lower in the strata containing small establishments.³

Osterman’s (1994) analysis of the NSE, like other studies of the adoption of innovative work practices, focuses on identifying the correlates of a collection of innovative work prac-

³See Osterman (1994) for more details of the survey design.

tices, measured with composite variables (for example, Pil and MacDuffie (1996) use a similar measure in their study of automobile assembly plants). Here, we focus on a single type of work practice: the use of self-directed work teams (SDWTs). Identifying factors associated with the use of teams is interesting in itself, and provides a focused and easily interpreted dependent variable for our study. The survey interviewers for the NSE defined SDWTs as teams in which “employees supervise their own work, they make their own decisions about pace and flow and occasionally the best way to get work done.” We thus operationalize SDWTs as a binary variable, equal to one if any employees in the establishment’s core job belong to such team, and equal to zero otherwise. The use of a binary variable has the additional advantage of allowing us to extend our discussion from ordinary least squares to logistic regression analysis.

We consider as independent variables several factors that might be associated with the use of SDWTs. Following Osterman (1994), we focus on structural characteristics and business strategies of the respondents. We consider seven factors: the number of employees in the establishment (*size*), whether or not unions are present in the establishment (*union*), whether or not the establishment is a branch of a larger establishment (*branch*), whether or not the establishment faces international competition (*internat*), an indicator of the extent to which the establishment competes by offering variety, high quality, and high levels of service relative to competing on the basis of low costs (*compete*), an indicator of the required skill level of the core jobs in the establishment (*skill*), and the self-reported values of the establishment’s managers (*value*).⁴ Each of these factors, with the exception of unionization, were significant correlates of the more general collection of innovative work practices in at least one of Osterman’s (1994) multivariate analyses of these data.

For analyses, we use only the 756 firms that have complete data for all variables of interest. We use this approach, which is different from the multiple imputation approach used in our analysis of the NOS, to illustrate adjustment of survey weights to handle missing data

⁴For complete variable definitions, see Osterman (1994).

when using survey-weighted analyses. The weights must be adjusted because they do not reflect probability of nonresponse. We assume the probability of providing an incomplete response depends solely on industry type (manufacturing versus non-manufacturing) and is constant within industry type. Essentially, this means the manufacturing firms with complete data are a simple random sample from the surveyed manufacturing firms, and likewise for the non-manufacturing firms. Under this nonresponse mechanism, the adjusted weights for manufacturing firms equal their basic weights multiplied by the inverse of the percentage of manufacturing firms with complete responses. This inflates the weights for manufacturing firms with complete responses so that they represent all the surveyed manufacturing firms. For non-manufacturing firms, the adjusted weights are defined similarly but use the percentage of non-manufacturing firms with complete responses. Further discussion of weight adjustments is provided later.

Analyzing Surveys With Complex Designs

Perhaps the simplest approach to analysis of any complex data set is to disregard the sampling design entirely and use methods of analysis that are appropriate for simple random samples. Implicitly, the approach rests on the claim that the sampling design is irrelevant for inferences once data are collected. This claim is simply false; the design typically does contain information that should be used in inferences. The NOS, for example, uses establishment size in the design, and size is often related to establishment work practices. More generally, in complex surveys, the sampling design frequently reflects characteristics of the population that are connected with the underlying relationships among the variables of interest. Disregarding the design increases the chance of missing such population structure, resulting in mis-specified models and invalid conclusions. In fact, as we discuss later, both design-based (survey-weighted) and model-based (unweighted) approaches are formally valid only when the design has been taken into account.

The best means of accounting for complex survey designs in analytical models is historically controversial (Brewer, 1973; Hansen *et al.*, 1983; Royall, 1983). The debate pits design-based versus model-based statistical methods. Because design-based methods are not tied to parametric models, they are robust to the veracity of certain assumptions. On the other hand, model-based methods are more flexible and, when the model assumptions are reasonable, more powerful than design-based methods. In what follows, we describe each of these methods of accounting for the survey design. Using our two data sets, we then illustrate and compare these different approaches to incorporating the design in analytical models.

Design-based Approaches

In the design-based approach to analytical inference, model parameters are treated like population quantities. They are estimated with survey-weighted averages of the sampled data, much like the Horvitz and Thompson (1952) estimators of population means. For example, consider the regression coefficients of a linear regression model that relates outcome y to predictors x . Let y_i be the outcome and x_i be the vector of predictors for the i th unit in the population, where $i = 1 \dots N$. In the design-based approach, we let the object of inference be the population regression coefficient, B , which equals

$$B = \left(\sum_{i=1}^N x_i x_i^t \right)^{-1} \sum_{i=1}^N x_i y_i. \quad (1)$$

To estimate B from the sample of units, S , we use a weighted average with each unit's weight, w_i , determined from the survey design:

$$b = \left(\sum_{i \in S} w_i x_i x_i^t \right)^{-1} \sum_{i \in S} w_i x_i y_i. \quad (2)$$

Similar weight-based estimators have been developed for logistic regression models (Chambless and Boyle, 1985), generalized linear models (Nordberg, 1989), and proportional hazard

models (Binder, 1992). Variances of these estimators can be estimated with Taylor series linearizations and resampling techniques (Wolter, 1985); variance estimators based on simple random samples typically are biased (Binder, 1983). Software packages that can implement these methods include SAS 8.0, Stata, WesVarPC, and Sudaan.⁵

Model-based Approaches

Model-based approaches do not use survey weights for estimation. Rather, model parameters are estimated using standard likelihood-based or Bayesian methods, for example ordinary least squares. This does not imply that sampling designs are irrelevant in model-based approaches. Actually, when relevant features of the sampling design are not accounted for, likelihood-based or Bayesian inferences are not statistically valid; that is, the inferences are biased and have poor statistical properties (Rubin, 1976; Sugden and Smith, 1984; Pfeffermann, 1993; Gelman *et al.*, 1995).

There is a simple principle of accounting for sampling designs in models: include independent variables that correspond to features of the sampling design. When this is done successfully, all information supplied by the survey design and survey weights is captured in the model. A mathematical justification of this claim, referred to as ignorability of the sampling design, can be found in Gelman *et al.* (1995, Chapter 7).⁶

The principle of ignorability does not lead directly to precise prescriptions for modeling relationships between outcomes y and given design variables z . The variety of data structures is too wide to make such precision possible. Particular sampling designs, however, do suggest certain structure. Consider three of the most popular sampling designs: probability proportional to size sampling, stratified sampling, and multi-stage sampling. For each design, there are different approaches to capturing possible underlying structure in the data.

⁵See Cohen (1997) and Zaslavsky (2004) for reviews of software for variance estimation in complex surveys.

⁶When the probabilities of being sampled are based on the dependent variable, it is impossible to account fully for the design by including appropriate independent variables. In this case, only the design-based (survey-weighted) approach can account for the survey design (Winship and Radbill, 1994). This case is atypical in analyses of surveys of work practices, since the variables on which survey designs usually are based—such as size, sector, and location—are not often selected as dependent variables for analysis.

- *Probability proportional to size sampling:* The design variable is a size measure, z , known for units in the sample. For example, if researchers use samples of individuals to generate a sample of workplaces, as in the National Organizations Survey, the probability that a given workplace is sampled is proportional to the number of employees in that workplace. For ignorability to hold, we need to include some function of z as a covariate in the model. The relationship between y and z may be nonlinear, so that polynomial terms or transformations of z may be needed. The model also may need terms for interactions between z and other covariates.
- *Stratified sampling:* The design variables are indicator variables for each of the strata. Two common forms of sampling in surveys of workplaces are stratification by industry and by size. The Census Bureau's National Employer Survey (Cappelli, 2001), for example, oversamples large manufacturing establishments. For ignorability to hold, we need to include some function of the stratum indicators in the model. This function might contain terms for separate intercepts, terms for interactions between the stratum indicators and other covariates, or terms for unequal variances across strata. Accounting for a design with industry stratification, for example, suggests inclusion of terms such as dummy variables for industries, and possibly interaction terms between industry dummies and other variables in the model.

When data are scarce in some strata, multi-level models (Bryk and Raudenbush, 1992; Gelman *et al.*, 1995) can be used to borrow strength across strata to improve fit. When strata boundaries are a function of a continuous covariate, such as a measure of size, it may be sufficient to include a function of this covariate instead of separate indicators for each stratum.

- *Multi-stage sampling:* Multi-stage sampling is increasingly common in workplace surveys that seek to learn about individuals within establishments as well as characteristics of the establishments themselves. As an example, consider first sampling establish-

ments from industry strata, then randomly sampling employees within the selected establishments. The design variables in multi-stage samples include all variables used at each level of sampling, such as stratum indicators (e.g., for industries), cluster indicators (e.g., for establishments), and any size measures. For ignorability to hold, the model should include some function of these variables. In our example, the design can be modeled by including indicator variables for the establishments; industry indicators are not needed because they are collinear with the establishment indicators. When modeling the design requires many indicator variables, inferences can be sharpened by relating these variables' coefficients with multi-level models; see Bryk and Raudenbush (1992) for several examples. As an alternative to indicator variables, analysts can include covariates that serve as proxies for aspects of the design. For example, when establishments are related to an outcome primarily through their size, analysts can include some function of size rather than indicator variables for all establishments.

Initial Analyses of the Data Sets

We now make a first pass at the analyses of the NOS and NSE, using the design-based (survey-weighted) and model-based (unweighted) approaches. For the NOS, we seek to predict *FILMScore* as a function of the independent variables described previously using linear regressions with normally distributed errors. Establishments are selected with probability proportional to *size*. Hence, following the guidelines for probability proportional to size designs, we should include some function of *size* in the regression. Of course, we would do this naturally because we are interested substantively in *size* as a covariate; ignorability provides a statistical rationale for doing so.

Ignorability does not tell us exactly how to specify the function of *size* in the model. We thus start with a straight-forward model: $\log(\textit{size})$ has a linear relationship with *FILMScore*, given the other covariates in the model. We use $\log(\textit{size})$ rather than *size* to reduce the influence of very large establishments on the values of the estimated regression coefficients.

The resulting estimated coefficients and standard errors for the model-based and design-based approaches are displayed in the first two columns of numbers in Table 1. Our first pass analysis reveals substantial differences between the model-based and design-based results. As examples, the unweighted and weighted coefficients for *formal*, *public*, *multisite*, and $\log(\textit{size})$ differ by factors of almost two or more, and are roughly two or more weighted standard errors apart; the hypothesis tests for the coefficients of *complex* and *public* have very different p-values, suggesting different conclusions about the importance of these variables; and, all standard errors for the weighted regressions are larger than those for the unweighted ones. Clearly, the choice of design-based or model-based approaches matters in the NOS.

The story for the analyses of the NSE is similar. In this case, the model of interest is a logistic regression of *SDWT* on the relevant independent variables. Because the NSE stratified on industry type and firm size, to follow the principle of ignorability we need to include industry type and size in the logistic regression. As a first pass, we adopt again a straight-forward modeling strategy: include separate indicator variables for each of the eleven industries represented in the sample. Manufacturing establishments make up one group, and non-manufacturing establishments are divided into ten separate industries. Thus we need ten indicators to identify the model (the omitted indicator for the reference category corresponds to the manufacturing industry). We also include a linear function of size. The choice to model size as a linear function, rather than as a series of indicator variables for the size strata, is motivated primarily from parsimony and ease of interpretation. Ignorability only requires that we control for the design variables: how we do so is up to us.

The resulting estimated coefficients and standard errors for the model-based and design-based approaches are displayed in the first two columns of numbers in Table 2. The coefficients of the 10 industry indicators are not displayed to save space. Once again, we see substantial differences in the approaches. In particular, the estimated coefficients for *union*, *compete*, *values*, and *size* are roughly two weighted standard errors apart in the weighted and unweighted models. Also, the p-values differ substantially between the two models for

almost every covariate in the model. The choice between weighted and unweighted models again has important consequences for inferences.

Comparisons of Statistical Properties of the Approaches

It is disconcerting that the model-based and design-based results for both the NOS and NSE differ so widely. Which approach yields better answers? In fact, ultimately we use neither the weighted nor unweighted estimates from the initial analyses, because we improve the fit of the models by altering the specification of the independent variables. We can use the results, however, as motivation for a general discussion on the advantages of and disadvantages of the design-based and model-based approaches, which are summarized in Table 3.

Design-based approaches have two main virtues. First and foremost, population quantities like B in Equation 1 exist independently of statistical models; hence, these quantities can be estimated using survey-weighted estimators that do not rely on assumptions about the probability distributions of the data. This virtue is a strong reason for some researchers' preference for the design-based approach. Second, and perhaps less important, weighted estimates allow approximately unbiased estimation of the coefficients of the independent variables in the regression model, even when other relevant independent variables are excluded from the model (Binder, 1983; Kott, 1991). This is because design-based estimators are asymptotically unbiased for the parameter values that would be obtained if the (possibly misspecified) regression model was fit using data from the entire population.

These virtues are limited. For practical purposes, the assumptions of the underlying analytic model must hold approximately; otherwise, the parameters of the model have no meaningful interpretation (Pfeffermann, 1993). An example illustrates this point. Suppose univariate x has different effects on univariate y in each of two strata. A researcher might choose to fit a model with different intercepts but a common slope for each strata, using weighted estimation of the two intercepts and one slope. The weighted estimate of the slope will be unbiased for the population quantity corresponding to the common slope. The

common slope, however, does not correctly describe the population: the interpretation of the estimate is practically meaningless.

Design-based approaches to analytical inference have further limitations. Resulting inferences are specific to the finite population under study, so that generalizations to different populations are not readily made (Kalton, 1983). Weighted estimators require large samples for confidence intervals and hypothesis tests to be valid (Pfeffermann, 1993). Model diagnostics for weighted analyses are not available on most standard software (e.g., SAS and Stata), so that the reasonableness of model assumptions needed to interpret the parameters are not easily checked.

Model-based approaches have several advantages over design-based approaches. Inference can be obtained from standard software and estimation routines, so that special techniques for weighted estimation are not needed. Most software packages, including SAS and Stata, permit easy examination of diagnostics for model-based approaches, but they do not have routines for diagnostics for design-based approaches. Importantly, constructing models to satisfy the conditions of ignorability encourages exploration of relationships between the outcome and the design variables, which can lead to more reliable modeling of relationships in the population.

In addition to these advantages, model-based estimators of analytic parameters typically have smaller standard errors than design-based estimators (Pfeffermann, 1993). This is because model-based parameter estimates, including those for linear and logistic regression, are usually obtained by maximum likelihood estimation, which has some optimal theoretical, statistical properties. In a rough sense, the unweighted estimates have smaller variances because they do not include the variation in the magnitude of the weights. When the weights vary widely, such as is the case in many complex samples, this additional variation in weights can be substantial, as is reflected in the larger standard errors for the weighted estimates as compared to the unweighted ones in the initial analyses of the NOS and NSE.

The main drawback to model-based approaches is that they can lead to unreliable con-

clusions when the models do not fit the data, though this is true of mis-specified models in any case. The model-based approach is arguably more sensitive to mis-specification, since the design-based approach does not rely on parametric assumptions. Thus, whatever the nature of the complexity of the design, it is crucial that models based on complex surveys be checked thoroughly for adequate fit to the data. Standard methods for simple random samples—such as those described by Belsley *et al.* (1980) for linear regression, by Hosmer and Lemeshow (2000) for logistic regression, and by Rubin (1984) and Bryk and Raudenbush (1992) for multi-level models—can be applied to data collected with complex samples.

Models estimated from complex surveys can be checked further by comparing the unweighted and survey weighted estimates of model parameters, as suggested by DuMouchel and Duncan (1983). When the proposed model fits the data well, the unweighted and weighted estimates should be similar, since both estimators are consistent (Pfeffermann, 1993). On the other hand, these estimates could differ substantially when the model does not fit well, so that re-specification of the model may be necessary. Judgments about these differences can be formalized with the hypothesis tests of DuMouchel and Duncan (1983) and Nordberg (1989), who suggest fitting augmented models that include (i) main effects for all independent variables, x , (ii) a main effect for the survey weights, w , and (iii) all two-way interactions between w and the variables in x . When the original model is correctly specified, w and its interactions with x should not be statistically significant predictors. When interactions with particular variables in x are statistically significant, the weights contain information that is not reflected in the independent variables, and additional modeling involving these x may improve model fit. Using both weighted and unweighted analyses also has been recommended by Lohr and Liu (1994), Winship and Radbill (1994), and Korn and Graubard (1995).

Finally, both the design-based and model-based approaches described here account only for the information in the sampling design and in the observed variables. There may be unobserved variables that strongly relate to the dependent variable of interest. If so, using

the weighted or unweighted models presented here will not eliminate the bias due to omitted variables. Different statistical methods may be required, such as instrumental variables (Heckman and Robb, 1985; Angrist *et al.*, 1996), selection models (Heckman, 1979, 1990), and sensitivity analyses in conjunction with propensity score matching methods (Rosenbaum and Rubin, 1983, 1984; Rosenbaum, 2002).

Improving the Analyses of the Correlates of Work Practices

With the above perspective on the two approaches, we now return to each of our examples to see if we can improve the models. We consider what effects our improved analyses have on the discrepancies between the weighted and the unweighted results. We then consider the implications of our improved analyses for the substantive conclusions one could draw about the correlates of internal labor markets and self-directed work teams.

The National Organizations Study

Because the weighted and unweighted coefficients of $\log(size)$ differ greatly, and $size$ is a design variable, we first consider whether the simple linear function of $\log(size)$ is appropriate. Graphical analyses suggest that the positive relationship between size and the use of firm internal labor markets could have a non-linear functional form. This misspecification of the relationship with size was confirmed using the significance tests based on the augmented model of DuMouchel and Duncan (1983) and Nordberg (1989), i.e., the model that includes all the main effects of the independent variables and their interactions with the survey weights. The p-value for the interaction involving $\log(size)$, shown in the first column in Table 4, is significant. This suggests we should re-specify how this variable enters the model.

Based on these diagnostics, we added $\log(size)^2$ to the original model. We tried other non-linear functions of $size$, but $\log(size)^2$ appeared to fit the data best. The results of the unweighted and weighted models including $\log(size)^2$ are displayed in the last two columns

of Table 1. The coefficient for $\log(\textit{size})^2$ is significant and negative, indicating that establishments' tendencies to adopt FILMs level out as establishment size increases.

The weighted and unweighted estimates for the coefficients of the terms involving *size* are now nearly identical in the final model. Adding the $\log(\textit{size})^2$ also reduces the difference observed in the initial model between the estimated coefficients for *formal*. However, the differences observed in the initial model for *public* and *multisite* remain. To assess these differences formally, we again fit an augmented model that includes all two-way interactions of the independent variables with the survey weights. The p-values of the corresponding significance tests for all the interactions are reported in the second column of Table 4. For most interactions, including the one for *formal*, the p-values are large. The p-values are not large for *public* and *multisite*, suggesting that we may need to modify the functions for these two variables in the model. We fit various interactions involving these two variables, but we did not find convincing evidence favoring the complicated interaction models over the simpler main effects model. It may well be that the final model is not correctly specified with regard to *public* and *multisite*, but we were not able to improve it.

The choice of the model-based or design-based approach for analysis of the NOS data continues to have substantive implications for conclusions about the correlates of FILMs. The unweighted model, for example, suggests that the extent of FILMs does not differ between public and private establishments, controlling for other characteristics. It also shows that environmental complexity (*complex*) is positively associated with *FILMscore*. In contrast, the weighted model suggests that private establishments are more likely to adopt FILMs than public establishments, and it does not identify the relationship between environmental complexity and FILMs at conventional levels of statistical significance, despite a weighted coefficient of nearly identical magnitude to the unweighted coefficient. While $\log(\textit{size})^2$ is highly significant in the unweighted model, the same term in the weighted model is not statistically significant, leading to different conclusions about the declining effect of firm size depending on the approach to handling the design. As anticipated, the standard errors of

the coefficients for the weighted regression are generally larger than those for the unweighted regression.

Since the residuals do not suggest violations of the linear model assumptions, we select the unweighted regression for interpretations to take advantage of the smaller standard errors. This model is very similar to the one employed by Kalleberg *et al.* (1996a), who modeled the relationship between size and FILMscore with a quadratic term in the logarithm of size and presented results of unweighted OLS regressions.

The NOS example illustrates an important consideration when comparing unweighted and weighted estimates as a model diagnostic. The comparison should be viewed primarily as another tool for diagnostics and model construction. Analysts need not feel obligated to fit certain interactions, or declare models inadequate, should there remain differences in the unweighted and weighted estimates. Rather, substantially different estimates compel the analyst to explore the possibility that modifying the model can improve the fit. Our concerns with accounting for the design thoroughly led us to include more than just simple functions of size or $\log(\text{size})$, and therefore to uncover the complex effect of size on the adoption of FILMs.

The National Survey of Establishments

For the NSE, the unweighted and weighted estimated coefficients for the initial model differ markedly for some variables, particularly *union*, *compete*, *values* and *size*. To check these differences formally, we use the significance tests based on the augmented model of DuMouchel and Duncan (1983) and Nordberg (1989). The p-values for the augmented model are shown in the third column of numbers of Table 2. The small p-values in the augmented model for *union*, *compete*, *values* and *size* confirm that we should re-examine the specifications of these variables.

We begin the model adjustment by focusing on *size*. Plots of *size* versus SDWT show that the relationship follows a pattern resembling an italicized, lowercase letter j; that is,

for small firms, increased firm size is associated with decreased propensity to use SDWTs. For medium and large firms, in contrast, increased firm size is associated with increased propensity to have SDWTs.⁷ We capture this pattern by including both *size* and $1/size$ in the model.

We next examine whether there are relationships among *size* and the other covariates or among industry and the other covariates. Fitting interactions of *size* or $1/size$ with each of the covariates does not yield significant improvement in the model; the change-in-deviance test statistics (Hosmer and Lemeshow, 2000) associated with these interactions are not statistically significant. Fitting interactions of all the industry indicators with the covariates or *size* is not feasible because of the small sample sizes in some industries. However, the coefficients of the industry indicators are similar enough to suggest that a single contrast between manufacturing and non-manufacturing firms can suffice for analyses. Therefore, we fit interactions of this contrast, which we call *Non-manuf*, with *compete*, *union*, and *values*. The interactions with *compete* and *union* appear to be useful predictors; each change-in-deviance test results in a small p-value. The interaction with *values* is not statistically significant.⁸ For completeness, we also fit interactions of *Non-manuf* with the remaining covariates, finding the interaction with $1/size$ to be a useful predictor.

Unweighted and weighted estimates from the final model are shown in the final two columns of numbers in Table 2. All weighted coefficients are within weighted two standard errors of the corresponding unweighted coefficients. To check these differences formally, we again use the significance tests based on the augmented model of DuMouchel and Duncan

⁷To make this plot, we first sorted the observations from smallest to largest in *size* and placed the ordered observations into 25 equally-sized groups. We graphed the groups' percentages of firms with $SDWT=1$ versus the groups' average sizes.

⁸We explored other transformations of the values variable, but these did not improve results. Exploratory analysis suggested that managers reporting that it would not be appropriate for the establishment to accept responsibility for the well-being of employees (a score of 1, the lowest possible) were less likely to have SDWTs. Only 11 of the 756 establishments, however, gave this response. Such a small number of low-scoring respondents left us reluctant to replace the continuous values variable with an indicator variable for the lowest score. These 11 establishments had larger survey weights on average than other respondents, contributing to the differences between the weighted and unweighted coefficients. The discrepancy suggests that future research should examine more closely the relationship between managers' values and the use of teams, paying particular attention to measurement of these values.

(1983) and Nordberg (1989), in which we fit interactions between the independent variables and their survey weights. The p-values are large (all greater than .10) for all of these interactions. Once again the standard errors for the unweighted estimates were substantially smaller than those from the weighted estimates. Checks of influence diagnostics based on the methods of Hosmer and Lemeshow (2000) revealed no model inadequacies or overly influential points.

The choice of approach again affects conclusions from the final model. The model-based approach, for example, showed statistically significant effect main effects of size and branch status; larger firms and branches were more likely to have teams. The model-based approach also showed a statistically significant negative main effect for unionization, but the design-based approach did not. The values variable, on the other hand, was statistically significant using the design-based approach but not using the model-based approach.

In the end, we prefer the unweighted, rather than weighted, final model for interpretations. The unweighted model has smaller standard errors. Further, with this model (unlike the weighted model), we were able to perform standard diagnostic checks, and we found no violations of model assumptions in these checks. Remaining differences between the two models, such as those for the values variable, suggest factors requiring more research and better measurement.

Our analysis of the NSE illustrates how use of the model-based approach, supplemented with the guidance provided by weighted estimates as diagnostics, reveals substantively important relationships among the variables. Our model-building process here revealed differences in the correlates of the use of teams by sector. The positive relationship between the use of teams and competitive strategies based on non-cost factors held only for the non-manufacturing establishments in the sample. In contrast, the positive coefficient on the union interaction term suggests that the negative relationship between unionization and the use of teams held only for the manufacturing firms. The relationship between size and the use of teams also differed by sector: teams were straightforwardly associated with larger

establishments in manufacturing, while in other sectors, the relationship was more complex and could not be captured with a linear term. If we had not thoroughly accounted for the design (for example, if we had used either of the initial models), we would not have uncovered these relationships.

Missing data in complex surveys

Both the NOS and NSE were plagued by item nonresponse, i.e. when units participate in a survey but provide incomplete responses. Item nonresponse complicates analyses and must be dealt with carefully. One approach is to use only units with complete data for all variables in the data set, often called the complete-cases approach (Little and Rubin, 2002, Chapter 3). An alternative is to use only units with no missing values for the variables of interest, often called the available-cases approach (Little and Rubin, 2002, Chapter 3). In either case, the sampling weights as originally established do not incorporate the selection of respondents among the surveyed units, so that the weights no longer reflect the representativeness of the sample. Using these incorrect weights in design-based estimation can result in biased estimates (Little and Rubin, 2002).

We have argued that researchers can benefit from using weighted analyses as diagnostics. If they do so, they may choose from two solutions to the problem of incorrect weights resulting from missing data. One approach is to use multiple imputation (Rubin, 1987), filling in the missing data repeatedly with draws from an imputation model based on the observed data. We did this in our analysis of the NOS data. Multiple imputation allowed us to analyze the completed data sets using the original survey weights, since all units in the data were included in the analyses. Advice on multiple imputation can be found in Schafer (1997), Schafer and Graham (2002), and Reiter and Raghunathan (2003). A second approach is to adjust weights to reflect nonresponse. We used this approach with the NES data. General advice on re-weighting can be found in Kalton and Kasprzyk (1986) and Sarndal *et al.* (1992,

Chapter 15).

Although it is not the main focus of our paper, it is worth comparing multiple imputation and weighting adjustments. Multiple imputation uses all units in the data file, whereas available-cases approaches throw away information from units with partially completed responses. Thus, inferences from multiple imputation tend to be more efficient than inferences from re-weighted, available-cases analyses (Little and Rubin, 2002). A related drawback of available-cases analyses is that different specifications of models may use different units for estimation, making theoretical properties of resulting inferences nearly impossible to understand and practical comparisons of different models difficult. This is especially onerous for weighted analyses, as it may necessitate several re-weightings as different observations enter the models. Multiple imputation is not plagued by these complications, since analyses can be based on all units in the data file regardless of the model.

Conclusions

Our discussion and analyses suggest that survey design must be incorporated into multivariate analyses if conclusions from models are to be reliable. The approach taken to do so can affect conclusions substantially, as illustrated in our comparisons of design-based and model-based analyses of the NOS and NSE. Hence, it is important to consider carefully the comparative advantages of the two approaches.

To obtain valid inferences with model-based methods, one should include functions of the design variables as predictors in the model. When this is done correctly, model-based estimates of coefficients have an advantage over design-based ones: they have smaller standard errors. In our analysis of the correlates of firm-level internal labor markets, for example, the standard error associated with the estimate for the effect of environmental complexity was so much larger in the weighted model than in the unweighted model that it suggested a different interpretation despite the nearly identical coefficient estimates. Similarly, the smaller

standard errors in the unweighted estimates of the union and branch effects on SDWTs led to different substantive conclusions than the weighted model would have suggested. A second advantage of the model-based approach is the relative ease of model building. Standard routines for (unweighted) model diagnostics are available in many statistical software packages; such routines typically do not exist for weighted models.

The advantages of the model-based approach disappear when the model fails to fit the data adequately. It is therefore crucial to check the model. Thorough use of standard diagnostics measures such as residuals provides some guidance for specifying the model. Comparisons of weighted and unweighted estimates can further improve the specification of models, and our analyses suggest that these comparisons should be an integral part of checking the fit of the posited analytic model. When the two sets of estimates are substantially different, re-specification of the model may be in order. In our consideration of the correlates of SDWTs, this diagnostic led us to identify interaction effects and j-shaped size functions that might otherwise have been difficult to find. We showed that several of the correlates of SDWTs differed between the manufacturing and service establishments in the sample. One particularly interesting product of this search was the discovery that unionization is negatively associated with teams, but only in manufacturing establishments.

Disregarding the sample design can have large consequences for inferences. It may come as some relief to researchers that actual survey implementation makes it very likely that the format of the design will reflect at least some of the underlying structure in the population. For example, establishment size is frequently one of the factors that affects sampling (and corresponding sample weights) in workplace surveys. Our discussion and examples suggest, however, that the relationship between size and work practices can be complex; simple linear functions of *size* or even $\log(\textit{size})$ may be inadequate. Careful attention to the design in analyses can help uncover such population structure, and thus lead researchers to better fitting models and more reliable conclusions.

References

- Althausen, Robert P. 1989. "Internal Labor Markets." *Annual Review of Sociology*, Vol. 15, pp. 143–61.
- Althausen, Robert P. and Arne L. Kalleberg. 1981. "Firms, Occupations, and the Structure of Labor Markets: A Conceptual Analysis." In Ivar Berg, ed., *Sociological Perspectives on Labor Markets*. New York: Academic Press, pp. 119–49.
- Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, Vol. 91, No. 433, pp. 444–55.
- Belsley, David A., Edwin Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.
- Binder, David A. 1983. "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review*, Vol. 51, No. 2, pp. 279–92.
- Binder, David A. 1992. "Fitting Cox's Proportional Hazards Model from Survey Data." *Biometrika*, Vol. 79, No. 1, pp. 139–47.
- Brewer, Kenneth R. W. and R. W. Mellor. 1973. "The Effect of Sample Structure on Analytical Surveys." *The Australian Journal of Statistics*, Vol. 15, No. 3, pp. 145–52.
- Bryk, Anthony S. and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications.
- Cappelli, Peter. 2001. "The National Employer Survey: Employer Data on Employment Practices." *Industrial Relations*, Vol. 40, No. 3 (July), pp. 635–47.
- Cappelli, Peter and David Neumark. 2001. "Do 'High-Performance' Work Practices Improve Establishment-Level Outcomes?" *Industrial and Labor Relations Review*, Vol. 54, No. 4, pp. 737–75.

- Chambless, Lloyd E. and Kerrie E. Boyle. 1985. "Maximum Likelihood Methods for Complex Sample Data: Logistic Regression and Discrete Proportional Hazards Models." *Communications in Statistics, Part A – Theory and Methods*, Vol. 14, No. 6, pp. 1377–92.
- Cochran, William G. 1977. *Sampling Techniques (3rd Ed)*. New York: John Wiley & Sons.
- Cohen, Steven B. 1997. "An Evaluation of Alternative PC-Based Packages for the Analysis of Complex Survey Data." *The American Statistician*, Vol. 51, No. 3, pp. 285–92.
- Doeringer, Peter B. and Michael J. Piore. 1971. *Internal Labor Markets and Manpower Analysis*. Lexington, MA: Heath.
- DuMouchel, William H. and Greg J. Duncan. 1983. "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples." *Journal of the American Statistical Association*, Vol. 78, No. 383, pp. 535–43.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. New York: Chapman & Hall.
- Gerhart, Barry, Patrick M. Wright, Gary C. McMahan, and Scott A. Snell. 2000. "Measurement Error in Research on Human Resources and Firm Performance: How Much Error Is There and How Does It Influence Effect Size Estimates?" *Personnel Psychology*, Vol. 53, No. 4 (Winter) pp. 803–34.
- Gramm, Cynthia L. and John F. Schnell. 2001. "The Use of Flexible Staffing Arrangements in Core Production Jobs." *Industrial and Labor Relations Review*, Vol. 54, No. 2, pp. 245–58.
- Hansen, Morris H., William G. Madow, and Benjamin J. Tepping. 1983. "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys." *Journal of the American Statistical Association*, Vol. 78, No. 388, pp. 776–93.

- Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica*, Vol. 47, No. 1, pp. 153–61.
- Heckman, James J. and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions: An Overview." *Journal of Econometrics*, Vol. 30 (October–November), pp. 239–67.
- Heckman, James J. 1990. "Varieties of Selection Bias." *The American Economic Review*, Vol. 80, No. 3 (June), pp. 313–18.
- Horvitz, Daniel G. and D. J. Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association*, Vol. 47, No. 260, pp. 663–85.
- Hosmer, David W. and Stanley Lemeshow. 2000. *Applied Logistic Regression, Second Edition*. New York: John Wiley & Sons.
- Hunter, Larry W. and Frits Pil. 1995. "How Do You Survey Firms?" *Proceedings of the Forty-Seventh Annual Meetings*. Madison, Wisconsin: Industrial Relations Research Association, pp. 152–62.
- Huselid, Mark A. 1995. "The Impact of Human Resource Management Practices on Turnover, Productivity, and Corporate Financial Performance." *Academy of Management Journal*, Vol. 38, No. 3 (June) pp. 635–72.
- Kalleberg, Arne L., David L. Knoke, Peter V. Marsden, and Joe L. Spaeth. 1996a. "Formalizing the Employment Relation." In Arne L. Kalleberg, David L. Knoke, Peter V. Marsden, and Joe L. Spaeth, eds., *Organizations in America: Analyzing Their Structures and Human Resource Practices*. Thousand Oaks, CA: Sage Publications, pp. 87–112.

- Kalleberg, Arne L., Peter V. Marsden, David L. Knoke, and Joe L. Spaeth. 1996b. *Organizations in America. Analyzing Their Structures and Human Resource Practices*. Newbury Park, CA: Sage Publications.
- Kalton, Graham. 1983. "Models in the Practice of Survey Sampling." *International Statistical Review*, Vol. 51, No. 1, pp. 175–88.
- Kalton, Graham and Daniel Kasprzyk. 1986. "The Treatment of Missing Survey Data." *Survey Methodology*, Vol. 12, No. 1, pp. 1–16.
- Korn, Edward L. and Barry I. Graubard. 1995. "Examples of Differing Weighted and Unweighted Estimates from a Sample Survey." *The American Statistician*, Vol. 49, No. 3, pp. 291–295.
- Kott, Phillip S. 1991. "A Model-Based Look at Linear Regression with Survey Data." *The American Statistician*, Vol. 45, No. 2, pp. 107–12.
- Little, Roderick J. A. and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data, Second Edition*. New York: John Wiley & Sons.
- Lohr, Sharon L. and Joanna Liu. 1994. "A Comparison of Weighted and Unweighted Analyses in the National Crime Victimization Survey." *Journal of Quantitative Criminology*, Vol. 10, No. 4, pp. 343–60.
- Lynch, Lisa M. and Sandra E. Black. 1998. "Beyond the Incidence of Employer-Provided Training." *Industrial and Labor Relations Review*, Vol. 52, No. 1, pp. 64–81.
- MacDuffie, John P. 1995. "Human Resource Bundles and Manufacturing Performance : Organizational Logic and Flexible Production Systems in the World Auto Industry." *Industrial and Labor Relations Review*, Vol. 48, No. 2, pp. 199–221.
- Nordberg, Lennart. 1989. "Generalized Linear Modeling of Sample Survey Data." *Journal of Official Statistics*, Vol. 5, No. 3, pp. 223–39.

- Osterman, Paul. 1984. "Introduction: The Nature and Importance of Internal Labor Markets." In Paul Osterman, ed., *Internal Labor Markets*. Cambridge, MA: MIT Press.
- Osterman, Paul. 1994. "How Common is Workplace Transformation and Who Adopts It?" *Industrial and Labor Relations Review*, Vol. 47, No. 2, pp. 175–88.
- Osterman, Paul. 2000. "Work Reorganization in an Era of Restructuring: Trends in Diffusion and Effects on Employee Welfare." *Industrial and Labor Relations Review*, Vol. 53, No. 2, pp. 179–96.
- Pfeffer, Jeffrey and Yinon Cohen. 1984. "Determinants of Internal Labor Markets in Organizations." *Administrative Science Quarterly*, Vol. 29, No. 4 (December), pp. 550–72.
- Pfeffermann, Danny. 1993. "The Role of Sampling Weights When Modeling Survey Data." *International Statistical Review*, Vol. 61, No. 2, pp. 317–37.
- Pil, Frits K. and John Paul MacDuffie. 1996. "The Adoption of High-Involvement Work Practices." *Industrial Relations*, Vol. 35, No. 3, pp. 423–55.
- Reiter, Jerome P. and Trivellore E. Raghunathan. 2003. "Multiple Imputation for Missing Data in Surveys with Complex Designs." Unpublished paper, Duke University.
- Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer-Verlag.
- Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, Vol. 70, No. 1, pp. 41–55.
- Rosenbaum, Paul R. and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association*, Vol. 79, No. 387, pp. 516–24.
- Royall, Richard M. 1983. "Comments on An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys." *Journal of the American Statistical Association*, Vol. 78, No. 384, pp. 794–96.

- Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika*, Vol. 63, No. 3, pp. 581–590.
- Rubin, Donald B. 1984. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician." *The Annals of Statistics*, Vol. 12, pp. 1151–72.
- Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Sarndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall.
- Schafer, Joseph L. and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods*, Vol. 7, No. 2, pp. 147–77.
- Spaeth, Joe L. and Diane P. O'Rourke. 1996. "Design of the National Organizations Study." In Arne L. Kalleberg, David L. Knoke, Peter V. Marsden, and Joe L. Spaeth, eds., *Organizations in America: Analyzing Their Structures and Human Resource Practices*. Thousand Oaks, CA: Sage Publications, pp. 23–39.
- Sugden, R. A. and T. M. F. Smith. 1984. "Ignorable and Informative Designs in Survey Sampling Inference." *Biometrika*, Vol. 71, No. 3, pp. 495–506.
- Winship, Christopher and Larry Radbill. 1994. "Sampling Weights and Regression Analysis." *Sociological Methods & Research*, Vol. 23, No. 2, pp. 230–57.
- Wolter, Kirk M. 1985. *Introduction to Variance Estimation*. New York: Springer-Verlag.
- Zaslavsky, Alan. 2004. Summary of survey analysis software. *American Statistical Association Survey Research Methods Section Webpage*. Retrieved January 27, 2004 from the World Wide Web <http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html>.

Table 1. OLS Regression Results for the Determinants of Firm Internal Labor Markets.

| Variable | <i>Initial Model</i> | | <i>Final Model</i> | |
|-----------------------|----------------------|---------------------|--------------------|---------------------|
| | <i>Model-based</i> | <i>Design-based</i> | <i>Model-based</i> | <i>Design-based</i> |
| <i>Intercept</i> | .73*** (.18) | .30 (.23) | .52*** (.19) | .22 (.24) |
| <i>lnlev</i> | .17*** (.04) | .15** (.07) | .13*** (.04) | .14* (.07) |
| <i>formal</i> | .70*** (.10) | .34** (.15) | .52*** (.11) | .33** (.15) |
| <i>depts</i> | -.14 (.12) | -.24 (.26) | .00 (.12) | -.19 (.27) |
| <i>decent</i> | .01 (.03) | .04 (.06) | .01 (.03) | .05 (.06) |
| <i>service</i> | -.02 (.08) | .01 (.09) | .01 (.08) | .01 (.09) |
| <i>prodserv</i> | .23** (.10) | .25* (.14) | .24** (.10) | .24* (.15) |
| <i>scope</i> | .02 (.02) | .03 (.04) | .01 (.02) | .03 (.04) |
| <i>lnage</i> | .04* (.02) | -.01 (.03) | .04 (.02) | -.01 (.03) |
| <i>eeprob</i> | .03 (.04) | .07 (.06) | .02 (.04) | .07 (.06) |
| <i>complex</i> | .19** (.09) | .21 (.13) | .19** (.09) | .20 (.13) |
| <i>public</i> | -.08 (.07) | -.37** (.15) | -.07 (.07) | -.36** (.15) |
| <i>nonprofit</i> | -.09 (.10) | .04 (.22) | -.08 (.09) | .04 (.22) |
| <i>instn</i> | -.02 (.02) | -.01 (.03) | -.03 (.02) | -.01 (.03) |
| <i>union</i> | -.01 (.04) | .12 (.12) | .01 (.04) | .14 (.12) |
| <i>multisite</i> | .25*** (.06) | .52*** (.14) | .26*** (.06) | .53*** (.14) |
| $\log(\text{size})$ | .15*** (.02) | .28*** (.05) | .36*** (.05) | .37*** (.11) |
| $\log(\text{size})^2$ | | | -.02*** (.01) | -.02 (.02) |
| Sample Size | 688 | 688 | 688 | 688 |
| R^2 Statistic | .61 | .63 | .62 | .63 |

Data come from the National Organizations Survey. Standard errors are in parentheses.

* indicates $.05 < p\text{-value} \leq .10$; ** indicates $.01 < p\text{-value} \leq .05$; *** indicates $p\text{-value} \leq .01$.

Table 2. Logistic Regression Results for the Determinants of Self-Directed Work Teams.

| Variable | <i>Augmented Model</i> | | | | | |
|------------------------------------|---------------------------------|---------------------|----------------------|--------------------|---------------------|-----|
| | <i>P-values of Interactions</i> | | | | | |
| | <i>Initial Model</i> | | <i>Predictors in</i> | | <i>Final Model</i> | |
| | <i>Model-based</i> | <i>Design-based</i> | <i>Initial Model</i> | <i>Model-based</i> | <i>Design-based</i> | |
| <i>Intercept</i> | -0.70 (.38) | -1.80* (.63) | .16 | -0.82* (.43) | -1.94*** (.70) | |
| <i>union</i> | -.51*** (.18) | .03 (.37) | .10 | -.72*** (.22) | -.40 (.33) | |
| <i>branch</i> | .55*** (.18) | .39 (.30) | .71 | .59*** (.19) | .47 (.33) | |
| <i>internat</i> | .26 (.17) | .52* (.31) | .49 | .34* (.18) | .64** (.31) | |
| <i>compete</i> | -.01 (.06) | .24** (.12) | .04 | -.12 (.08) | -.07 (.13) | |
| <i>values</i> | .07 (.09) | .33** (.15) | .11 | .08 (.09) | .32** (.16) | |
| <i>skill</i> | .53*** (.17) | .58* (.31) | .75 | .53*** (.17) | .64** (.33) | |
| <i>size * 1000</i> | .15*** (.04) | -.29 (.22) | .04 | .19*** (.05) | .04 (.11) | |
| <i>1/size</i> | | | | 5.7 (21.3) | 9.9 (29.3) | |
| <i>Non-manuf × compete</i> | | | | .32** (.14) | .56** (.26) | |
| <i>Non-manuf × 1/size</i> | | | | 83.5*** (30.7) | 104.0** (46.4) | |
| <i>Non-manuf × union</i> | | | | .64* (.38) | .67 (.71) | |
| Sample Size | 756 | 756 | 756 | 756 | 756 | 756 |
| Pseudo ^a R ² | .07 | NA ^b | — | .09 | NA ^b | |

^a Pseudo $R^2 = 1 - (\text{loglikelihood for model})/(\text{loglikelihood for model with intercept only})$.

^b Stata 7.0 does not provide a pseudo R^2 for weighted logistic regression models.

Data come from National Survey of Establishments. Standard errors are in parentheses.

Models include estimates for industry indicators; these are available upon request.

* indicates $.05 < \text{p-value} \leq .10$; ** indicates $.01 < \text{p-value} \leq .05$; *** indicates $\text{p-value} \leq .01$.

Table 3. Summary of Comparison of Design-based and Model-based Approaches.

| | <i>Design-based</i> | <i>Model-based</i> |
|---|---------------------|-----------------------------|
| <i>Probability distributions needed for estimation?</i> | no | yes |
| <i>Specialized software needed to fit models?</i> | yes | no |
| <i>Standard errors</i> | larger | smaller |
| <i>Model diagnostics</i> | difficult | standard routines available |

Table 4. P-values for Augmented Models in Analyses of the
Determinants of Firm Internal Labor Markets.

| <i>Variable</i> | <i>P-values of Interactions of Weights and Predictors</i> | |
|---|---|--------------------------------|
| | <i>Initial Model^a</i> | <i>Final Model^b</i> |
| <i>weight</i> × <i>Intercept</i> | .78 | .38 |
| <i>weight</i> × <i>lnlev</i> | .85 | .76 |
| <i>weight</i> × <i>formal</i> | .51 | .43 |
| <i>weight</i> × <i>depts</i> | .59 | .76 |
| <i>weight</i> × <i>decent</i> | .68 | .71 |
| <i>weight</i> × <i>service</i> | .99 | .91 |
| <i>weight</i> × <i>prodserv</i> | .92 | .91 |
| <i>weight</i> × <i>scope</i> | .90 | .89 |
| <i>weight</i> × <i>lnage</i> | .75 | .80 |
| <i>weight</i> × <i>eeprob</i> | .68 | .68 |
| <i>weight</i> × <i>public</i> | .07 | .07 |
| <i>weight</i> × <i>complex</i> | .95 | .84 |
| <i>weight</i> × <i>nonprof</i> | .68 | .63 |
| <i>weight</i> × <i>instn</i> | .74 | .68 |
| <i>weight</i> × <i>union</i> | .56 | .55 |
| <i>weight</i> × <i>multisite</i> | .09 | .08 |
| <i>weight</i> × <i>log(size)</i> | .01 | .22 |
| <i>weight</i> × <i>log(size)</i> ² | | .07 |

^a P-values in Column 1 are associated with hypothesis tests for the interactions of the weights and the independent variables added to the model in Column 1 of Table 1.

^b P-values in Column 2 are associated with hypothesis tests for the interactions of the weights and the independent variables added to the model in Column 3 of Table 1.