8

COMPLEX SAMPLE DATA IN STRUCTURAL EQUATION MODELING

*Bengt O. Muthén**
Albert Satorra†

Large-scale surveys using complex sample designs are frequently carried out by government agencies. The statistical analysis technology available for such data is, however, limited in scope. This study investigates and further develops statistical methods that could be used in software for the analysis of data collected under complex sample designs. First, it identifies several recent methodological lines of inquiry which taken together provide a powerful and general statistical basis for a complex sample, structural equation modeling analysis. Second, it extends some of this research to new situations of interest. A Monte Carlo study that empirically evaluates these techniques on simulated data comparable to those in large-scale complex surveys demonstrates that they work well in practice. Due to the generality of the approaches, the methods cover not only continuous normal variables but also continuous non-normal variables and dichotomous variables. Two methods designed to take into account the complex sample structure were

Muthén's research was supported by the National Science Foundation grant SES-8821668 and by the National Institute on Alcohol Abuse and Alcoholism Small Business Innovation Contract No. ADM 281-90-004. Satorra's research was partially supported by the Spanish DGICYT grant PB91-0814. The authors would like to acknowledge helpful comments from Roderick Little and Linda Muthén and the research assistance of Ginger Nelson Goff and Kathleen Wisnicki.

*University of California, Los Angeles
†Universitat Pompeu Fabra, Barcelona

investigated in the Monte Carlo study. One method, termed aggregated analysis, computes the usual parameter estimates but adjusts standard errors and goodness-of-fit model testing. The other method, termed disaggregated analysis, includes a new set of parameters reflecting the complex sample structure. Both of the methods worked very well. The conventional method that ignores complex sampling worked poorly, supporting the need for development of special methods for complex survey data.

1. INTRODUCTION

Large-scale surveys using complex sample designs are frequently carried out by government agencies. For example, the National Center for Health Statistics produces an annual survey of the nation's civilian, noninstitutionalized population regarding basic health issues; the Bureau of Labor Statistics sponsors the National Longitudinal Surveys, which collect data on labor market experiences of several different subpopulations; and the National Center for Education Statistics sponsors educational assessments such as the National Education Longitudinal Study. The multivariate statistical analysis technology available for such data is, however, limited in scope.

This study investigates and further develops multivariate statistical methods for the analysis of data collected under complex sample designs. Recent methodological lines of inquiry for complex sample data are discussed as they pertain to structural equation modeling. The study extends some of this research to new situations of interest. Two methods designed to take into account complex sample structure are investigated in a Monte Carlo study. One method, termed aggregated analysis, computes the usual parameter estimates but adjusts standard errors and goodness-of-fit testing. The other method, termed disaggregated analysis, includes a new set of parameters reflecting the complex sample structure. Due to the generality of the approaches, the methods cover not only continuous normal variables but also continuous nonnormal variables and dichotomous variables.

To introduce the complex sampling issues typically found in large-scale surveys, the features of two such surveys will be described. The 1988 National Health Interview Survey (NHIS) included a large supplement concerning alcohol consumption patterns sponsored by

the National Institute on Alcohol Abuse and Alcoholism (NIAAA). Its features are typical of those found in large-scale, national surveys. The population is the civilian, noninstitutionalized U.S. population. The NHIS is a complex multistage probability sample. A total of 1,900 primary sampling units (PSUs) defined by counties are stratified into 52 self-representing (SR) and 73 non–self-representing (NSR) strata. The SR strata are PSUs with the largest populations and are included in the sample with probability one. From each NSR stratum, two PSUs are chosen, without replacement and proportional to size (Durbin 1967). Within each PSU, three substrata are used: housing units from the building permit frame, housing units in the area frame to be oversampled (the study oversamples black persons), and other housing units in the area frame. Within each substratum, households are grouped into clusters of secondary sampling units (SSUs), also called segments. For substratum one, four housing units are expected, while for the other two substrata, eight households are expected. Selection of SSUs is by systematic sampling with a random start. All households within a sampled SSU are targeted for an interview. For the NHIS88 supplement, a household member 18 years of age or older was randomly selected. The NHIS88 alcohol survey resulted in 43,809 interviews.

Another major survey, the National Longitudinal Study of the High School Class of 1972 (NLS), concerns educational data. In this survey, the population consists of persons who were twelfth-grade students in U.S. schools during 1971–72. The sample design for the base year drew two schools without replacement from each of 600 strata. Equal probability selection was used, with the major exception that schools in low-income areas and with a high proportion of minority enrollment were oversampled. In a second stage, 18 students in each school were selected by simple random sampling. Starting with the fifth follow-up, an unequal probability subsample of the participating students was chosen, with oversampling of certain targeted student groups.

In reviewing the complex sample features that must be taken into account in statistical analyses of such data, it is convenient to distinguish between issues related to stratification and unequal inclusion probabilities on the one hand and issues related to clustering on the other hand. In the NHIS88 example, the stratification and oversampling features cause the households to have different probabili-

ties of being selected. This is reflected in the basic weights, which are inverse to the probabilities of inclusion. In the NLS, the oversampling of certain student categories for follow-up is another important source of unequal weights. If these are not taken into account, biased parameter estimates may result unless sample selection is based on variables that are exogenous in the model.

Clustering is exemplified by the NHIS hierarchical data structure in which households are observed within segments and segments are observed within PSUs. Observations on sample units that share segment or PSU membership may be correlated, and if this is ignored the standard errors of the parameter estimates are usually underestimated. Correlated observations may be a particularly important issue when students are observed within classes and classes within schools, as in the NLS.

It is also important to make distinctions among analytic approaches to estimation and testing. Several such distinctions appear in the literature. These will be described along with the distinctions used in this study. Many find it convenient to distinguish design-based and model-based analyses. This usually refers to the statistical philosophy of inference (e.g., see Skinner, Holt, and Smith 1989, p. 17). Design-based (randomization theory) approaches have been developed mainly by sampling statisticians interested in estimating finite population quantities (such as the population total and the linear regression function for all members of the population), whereas model-based approaches are more in line with conventional statistical modeling where estimators are derived by assuming a certain (super-) population model, which is probably only approximately true for all members of the population (e.g., see Hansen, Madow, and Tepping 1983, and discussions of this paper). Design-based inference refers to the sampling distribution of repeated samples generated by the sampling design, whereas model-based inference refers to the sampling distribution generated by the model. Another distinction may be drawn between aggregated and disaggregated approaches to analysis (Skinner et al. 1989, p. 8). In an aggregated analysis, model parameters are defined without conditioning on the design variables, whereas in a disaggregated analysis they are defined conditionally.

This study concerns both design- and model-based issues and aggregated and disaggregated modeling. Typically, design-based analysis uses weights in parameter estimation and Taylor linearization or

sample reuse methods to compute standard errors of estimators so that these properly reflect the likely variation in the estimates due to repeated sampling. Disaggregated modeling includes design features explicitly in the model by using additional parameters—for example, allowing for model heterogeneity across strata and clusters. Examples include the use of design variables as explanatory variables in regression, and inclusion of variance components corresponding to the various levels of clustering. This type of analysis is often model based.

One may claim that a design-based approach is less ambitious than a model-based approach. For example, one may view a design-based regression analysis as merely attempting to estimate the best overall regression model in a heterogeneous population, with the aim of obtaining test statistics that are valid in repeated sampling. In contrast, the model-based disaggregated approach may be viewed as an attempt not only to conduct correct tests but also to disentangle population heterogeneity by providing, for example, variance parameter estimates broken down into components for households, segments, and PSUs. Such approaches may, however, be more sensitive to model misspecification. It will be of interest to study whether design-based analysis features can be incorporated into model-based disaggregated analyses. More detailed examples will be given below in the discussion of disaggregated structural equation modeling.

The outline of the paper is as follows. Section 2 reviews previously developed design- and model-based methods for handling data from complex samples. Section 3 reviews conventional models and estimation procedures that do not take into account the sample design. Sections 4 and 5 describe the two principal statistical procedures for structural equation modeling with complex sample data, one being an aggregated approach, and the other a disaggregated approach. Section 6 presents the results of a Monte Carlo study that illustrates the promise of these approaches by comparison with a data analysis that ignores the sample design. Section 7 summarizes the discussion.

2. PRIOR APPROACHES TO COMPLEX SAMPLE DATA

2.1. *Univariate, Design-Based Methods*

To date, most statistical developments for complex samples have focused on univariate, design-based methods, placing special empha-

sis on taking sampling weights into account in the estimation of parameters and computing the proper standard errors of parameter estimates. Kish and Frankel (1974) pointed to three methods for computing standard errors that are now classic: Taylor expansion (linearization), balanced repeated replication (BRR), and jacknife repeated replication. Today, one may add bootstrap techniques to this list (e.g., see Rao and Wu 1988). For overviews of these methods, see Wolter (1985) and Rust (1985).

The Taylor expansion method (e.g., see Woodruff 1971) is applied to estimators such as linear regression with weighting. The probability weights are utilized in Horvitz-Thompson estimators (Cochran 1977, ch. 9A.7) for the sums of squares and cross-product matrices to provide estimates of the population regression coefficients, and a first-order Taylor expansion yields large-sample approximations to the standard errors of the estimates (e.g., see Shah, Holt, and Folsom 1977; Holt, Smith, and Winter 1980, Procedure 3). More recently, Binder (1983) gave a general method for obtaining estimates and standard errors for generalized linear models in a variety of common applications, including ordinary regression, logistic regression, and log-linear models for categorial data. A unifying approach to obtaining standard errors via Taylor linearization was given. Related quasi-maximum likelihood methods were discussed in Skinner et al. (1989) and McCullagh and Nelder (1989).

A general approach to estimation of logit models for weighted data was taken by Landis et al. (1987), who used a Taylor expansion to provide a covariance matrix for logits, and fitted models by generalized least squares with Wald tests. The BRR technique was also proposed for a general set of estimators. For categorical data, chi-square testing adjusted for complex sampling was also considered in Rao and Thomas (1988) and references therein.

Comparisons of these standard, design-based methods seem to indicate that the Taylor method gives slightly better results in the estimation of sampling variances, while BRR and other replication methods give better confidence interval coverage (e.g., see Kish and Frankel 1974; Rao and Wu 1985; see also Flyer, Rust, and Morganstein 1989). The differences are often not large, however, and both types of methods have been found quite acceptable in many situations (e.g., see Shah et al. 1977; Bean 1975; Wilson, 1989). Bean's study is important from the point of view of the NHIS, since

her conclusions were based on a large number of repeated samples from the 1969 U.S. Health Interview Survey data. Bean also found that tests and confidence intervals generated from either of these complex sample variance estimators could be well approximated by normal distribution theory.

2.2. *Univariate, Model-Based Methods*

Univariate, model-based methods generally deal with either unequal inclusion probabilities or clustering, but not both. Consider first those studies that take into account unequal inclusion probabilities. Estimation of finite population means was studied in Little (1983) and regression modeling in, for example, Holt, Smith, and Winter (1980), Nathan and Holt (1980), Pfeffermann and Holmes (1985), and Pfefferman and LaVange (1989). For regression models, these authors considered maximum likelihood estimation including design variables as covariates. They contrasted the performance of these techniques with the design-based approach of probability-weighting involving variance estimation by Taylor expansion. Limited simulations reported in these papers appear to indicate that modeling approaches can give considerably better mean square error performance than the probability-weighted approach, particularly because they yield considerably smaller variance estimates. When a model is incorrectly specified, however, this advantage may be lost.

An interesting example involving heteroscedastic residuals in model-based regression is given in Skinner et al. (1989, pp. 65–67). Little (1989) considered models with parameters that vary randomly across strata, leading to Bayesian estimation. This approach gives a modeling rationale for probability-weighted estimation.

Consider next model-based methods that take into account effects of clustering. A classic article in this area is by Scott and Smith (1969), who take a random parameter, Bayesian, approach. This in effect specifies a variance component model that explicitly models the correlations among observations within the clusters (see also Fuller and Battese 1973). More recently, Malec and Sedransk (1985) and Battese, Harter, and Fuller (1988) have suggested similar variance component models for samples with clustering (see also Scott and Holt 1982). A host of articles have dealt with related random effects modeling in the regression context, particularly for

longitudinal studies, including both continuous and categorical dependent variables; see Laird and Ware (1982) and, for recent overviews, Diggle, Liang and Zeger (1994), Longford (1993), and Rutter and Elashoff (1994). These approaches also pertain to hierarchically obtained data (e.g. members within a household), with both random intercepts and random slopes. They have recently become popular in educational research with students observed within classrooms and schools; see, for example, Bock (1989) and Bryk and Raudenbush (1992). Estimation of (super-) population parameters is most often carried out by maximum likelihood, while group- (cluster-) specific quantities are estimated by empirical Bayes methods. These developments do not take into account the unequal probabilities of selection usually encountered in complex surveys.

Longford (1989, 1993), provides an interesting application of variance component (or random coefficient regression) techniques to complex sampling, which is of direct relevance to this study. He considered data from the National Assessment of Educational Progress (NAEP), which has a multistage sample design with 32 strata, from each of which two PSUs were selected with replacement. Schools were sampled within each PSU, and students were sampled within schools. Minority groups were oversampled and sampling weights adjusted for nonresponse. Using simulated data with characteristics like those of the NAEP, Longford contrasted the performance of the design-based jackknife with maximum likelihood estimation of a three-level variance component model. Longford concluded that the computationally intensive jackknife procedure was outperformed by the variance component approach. Although estimated means were not appreciably different under the two procedures, means could be estimated with considerably more precision using the variance component approach.

2.3. *Multivariate Methods*

Methods for multivariate response models will now be discussed. Relatively little statistical research work has been carried out for multivariate analysis of data generated from complex samples. There are, however, at least two important areas in which multivariate methodology is emerging: log-linear modeling and structural equation modeling.

Complex sampling aspects of log-linear modeling of multivariate categorical data in frequency table form were discussed in, for example, Freeman et al. (1976); Landis et al. (1987); and Rao and Thomas (1988). These authors consider both unequal selection probabilities and clustering. A general weighted least squares (GLS) approach is used to estimate parameters and test hypotheses, where a set of sample statistics s (here a vector of proportions) is analyzed using a suitable weight matrix W. Using W as an approximation to the covariance matrix of s yields the approach of generalized least squares, and subsequently the Wald statistic. Landis et al. (1987) discussed Taylor-series approximations to W, and Rao and Thomas (1988) discussed sample reuse methods (jackknife, BRR). Alternative approaches include the Rao-Scott "generalized deff matrix" weighting, and, for hypothesis testing, Fay's jackknifed chi-square test and the Rao-Scott first- and second-order corrections to the simple random sample chi-square (see Rao and Thomas 1988).

Weighted least squares (WLS) estimators of this type are also considered in structural equation modeling where s contains the elements of the sample covariance matrix (e.g., see Jöreskog and Sörbom 1989; Bentler 1989; McDonald 1980; Muthén 1987). A common alternative is maximum likelihood estimation under normality assumptions. WLS estimation with a general weight matrix W has been considered, for example, in Browne (1982, 1984). An adaptation to take complex sampling into account when estimating s and W has recently been discussed in Skinner et al. (1989, ch. 3) and in Satorra (1992), but it does not appear to have been used in practice. In the current chapter's terms, this type of modeling may be described as aggregated.

Disaggregated, model-based, multivariate methodology analogous to the univariate case of variance component estimation and random effects regression is now emerging (see also Skinner et al. 1989, ch. 8). Maximum likelihood estimation of covariance structure models with latent variables, including random effects for hierarchical data, was considered in Goldstein and McDonald (1988), McDonald and Goldstein (1989), Lee (1990), Longford and Muthén (1990), and Muthén (1989*a*). Muthén (1990) considered maximum likelihood estimation of a factor analysis model for a two-level data structure with individuals observed within groups (see also Muthén 1991). This type of modeling can take into account correlations

across observations that result from clustering, extending the variance component approach to multivariate response models with latent variables. As pointed out in Muthén (1990), such an analysis appears to be computationally feasible, but to date the statistical developments have been very limited and no general software has been developed. Muthén (1994*a*) discusses multilevel covariance structure approaches using conventional structural equation modeling software, and gives an overview of work to date. Multivariate approaches will be further discussed in the following sections.

3. CONVENTIONAL STRUCTURAL EQUATION MODELS AND METHODOLOGY

In this section, standard approaches to structural equation modeling are described. This includes model specification, estimation schemes, standard errors of estimates, and goodness-of-fit statistics.

The statistical methodology outlined in this section provides a general framework which accommodates both aggregated and disaggregated modeling. Aggregated modeling may involve analysis of heterogeneous populations (e.g. see Muthén, 1989*a*) and issues related to this will be discussed below. The proposed methodology for calculation of standard errors and test statistics takes into account not only complex sampling but also nonnormality of variables. For this reason, the methodology is also preferable for simple random samples. First, the general model is described. This is followed by a discussion of estimation and testing under simple random sampling, assuming that observations are independently and identically distributed (iid).

3.1. *The model*

Consider a p—dimensional vector of observed variables y_i and an m—dimensional vector of latent variables η_i for observation unit i. Assume that for a certain population the following structural equation model holds:

$$y_i = \nu + \Lambda\eta_i + \epsilon_i, \tag{1}$$

$$\eta_i = \alpha + B\eta_i + \zeta_i, \tag{2}$$

where ν is a $p \times 1$ measurement intercept vector, Λ is a $p \times m$ matrix of measurement slopes, ϵ is a $p \times 1$ measurement error residual vector, α is an $m \times 1$ vector of structural intercepts, B is an $m \times m$ matrix of structural regression slopes with zero diagonal elements such that $(I - B)$ is not singular, and ζ is an $m \times 1$ vector of structural residuals. Equation (1) specifies the measurement part of the model and (2) specifies the structural regression part. For the special case of $\nu = 0$, $\Lambda = I$, and $\epsilon = 0$, a regression model is obtained, with coefficients estimated in B. For the special case of $\alpha = 0$ and $B = 0$, a factor analysis model is obtained with factor loadings estimated in Λ. The covariance structure for (1) and (2) covers essentially all structural equation models presented to date. For simplicity, mean structure models and multiple-group analyses will not be discussed here, but all techniques are directly generalizable to these situations (e.g., see Muthén 1983, 1984, 1987).

With usual assumptions, the covariance matrix of y is

$$\Sigma = \Lambda(I - B)^{-1}\Psi(I - B)^{-1\prime}\Lambda' + \Theta, \tag{3}$$

where Ψ is the $m \times m$ covariance matrix of ζ and Θ is the $p \times p$ covariance matrix of ϵ. The covariance structure model parameters are contained in the arrays Λ, B, Ψ and Θ. Let these parameters be assembled in the $q \times 1$ parameter vector κ. The model structure may then be expressed as $\Sigma(\kappa)$. In conventional structural equation modeling, κ is estimated by fitting $\Sigma(\kappa)$ to S, the covariance matrix for a sample of n observations on y.

3.2. *Estimation and Testing Under Iid Assumptions*

This section discusses inferential procedures that involve not only normal variables but also nonnormal variables. The distributions of the variables measured in surveys are frequently very skewed because they measure behaviors in which only a minority of the population is engaged. Moreover, from a statistical point of view, we show that techniques for nonnormal data are special cases of techniques for complex sample data.

Under the conventional assumption of iid, and normally distributed observations on y, the sample covariance matrix S contains the sufficient statistics for estimating κ. In this case, two common

fitting functions are normal theory maximum likelihood (NTML) and normal theory GLS (NTGLS),

$$F_{NTML} = \ln |\Sigma| + \text{trace}(\Sigma^{-1}S) - \ln |S| - p \tag{4}$$

$$F_{NTGLS} = \text{trace}[(\Sigma - S)S^{-1}]^2. \tag{5}$$

The expression for F_{NTGLS} is a special case of the general weighted least squares fitting function

$$F_{WLS} = (s - \sigma)'W^{-1}(s - \sigma) \tag{6}$$

where s and σ refer to the $p^* = (p(p + 1)/2)$-dimensional vectors of distinct elements of S and Σ, respectively. If in (6), W is taken as a consistent estimate of the asymptotic covariance matrix of s, then F_{WLS} is known as a generalized least-squares estimator or minimum chi-square analysis (Ferguson 1958; Fuller 1987, sect. 4.2; see also Satorra 1992). In the special case where y is multivariate normal, the asymptotic covariance matrix of s has a particularly simple structure, depending only on second-order moments,

$$W_{NT} = 2D^+(\Sigma \otimes \Sigma)D^{+\prime}, \tag{7}$$

where D^+ is an "elimination" matrix (see Magnus and Neudecker 1988) and $\otimes$ denotes the Kronecker product. A consistent estimator of (7) is obtained by replacing Σ with S. The use of W_{NT} in (6) leads to (5) (e.g., see Satorra 1992).

For arbitrary distributions, the asymptotic covariance matrix of s—say Γ—has elements

$$\gamma_{ijkl} = \sigma_{ijkl} - \sigma_{ij}\sigma_{kl} \tag{8}$$

(e.g., see Browne 1982), when these moments exist. Define the p^*-dimensional data vector d_i for observation i,

$$d_i \equiv \begin{pmatrix} (y_{i1} - \bar{y}_1)(y_{i1} - \bar{y}_1) \\ (y_{i2} - \bar{y}_2)(y_{i1} - \bar{y}_1) \\ (y_{i2} - \bar{y}_2)(y_{i2} - \bar{y}_2) \\ \vdots \\ (y_{ip} - \bar{y}_p)(y_{ip} - \bar{y}_p) \end{pmatrix} \tag{9}$$

where y_{iv} is the *ith* observation on variable $v (v = 1, 2, \ldots, p)$ and $\bar{y}_v$ is the corresponding mean. For sample size n, we thus have

$$(n-1)^{-1}\sum_{i=1}^{n} d_i = s \tag{10}$$

A consistent estimator of Γ is obtained as the sample covariance matrix of the d_i involving fourth-order moments (e.g., see Browne 1982, 1984; Chamberlain 1982; see also Satorra 1992)

$$\hat{\Gamma} \equiv (n-1)^{-1}\sum_{i}^{n} (d_i - \bar{d})(d_i - \bar{d})' \tag{11}$$

Taking $\hat{\Gamma}$ as W in the weighted least squares fitting function of (6) gives the asymptotically distribution free (ADF) estimator proposed by Browne (1982) for covariance structure analysis of nonnormal continuous variables.

Consider now standard errors of parameter estimates and tests of model fit. Define the $p^* \times q$ matrix of derivatives

$$\Delta = \partial\sigma(\kappa)/\partial\kappa. \tag{12}$$

Estimating κ with the weighted least squares fitting function of (6), it is well-known that a Taylor expansion gives the asymptotic covariance matrix

$$\text{acov}(\hat{\kappa}) = n^{-1}(\Delta' W^{-1}\Delta)^{-1}\Delta' W^{-1}\Gamma W^{-1}\Delta(\Delta' W^{-1}\Delta)^{-1} \tag{13}$$

(e.g., see Ferguson 1958; Browne 1984; Fuller 1987; and Satorra 1992). When $W = \hat{\Gamma}$, the above asymptotic variance matrix simplifies to

$$\text{acov}(\hat{\kappa}) = n^{-1}(\Delta' W^{-1}\Delta)^{-1} \tag{14}$$

Expression (14) is commonly used for both *NTGLS* and *ADF*. The same expression can be shown to hold for *NTML* (e.g., Satorra 1989). The above variance matrices can be consistently estimated by evaluating $\Delta = \Delta(\kappa)$ at $\hat{\kappa}$ and replacing Γ by its consistent estimate (11).

The expression in (13) shows that W need not be the same as $\hat{\Gamma}$. For example, W may be calculated via the computationally simple normal theory expression in (7) to give normal theory parameter estimates. Using the general $\hat{\Gamma}$ expression of (11) in (13) provides a proper covariance matrix for these estimates even under nonnormality. Such an approach is strongly preferable to ADF from a com-

putational point of view when the number of variables is large. When obtaining standard errors via (13), the large matrix $\hat{\Gamma}$ need not be inverted. In contrast, in the ADF approach $\hat{\Gamma}$ must be inverted even if standard errors are not required (see [6] and [14]).

Under normal theory, the conventional model test of fit of H_0 against an unrestricted covariance matrix is obtained as the likelihood ratio statistic $n\hat{F}$, where $\hat{F} = F(\hat{\kappa})$ is the value of the fitting function ([4] or [5]) at its maximum. For (6) a corresponding Wald statistic is obtained. This quantity is distributed as chi-square with $p^* - q$ degrees of freedom. Relaxing the restriction of normality, Browne (1984) gave a more general expression for a chi-square goodness-of-fit test. Consider the following quadratic form on the residuals (Browne 1984; Satorra and Bentler 1990, 1994; Satorra, 1990, 1992):

$$nT = (s - \hat{\sigma})' A(s - \hat{\sigma}) \tag{15}$$

where $\hat{\sigma} = \sigma(\hat{\kappa})$ and A is a consistent estimator of

$$W - W\Delta(\Delta' W\Delta)^{-1}\Delta' W = \Delta\perp(\Delta\perp' W^{-1}\Delta\perp)^{-1}\Delta\perp', \tag{16}$$

where $\Delta\perp$ is an orthogonal complement of Δ.

Robustness to nonnormality is obtained by using $W = \hat{\Gamma}^{-1}$ of (11). Satorra (1989) shows that this also holds when the optimum of F is obtained via NTML. A simpler, mean-corrected expression is the scaled chi-square, $n\hat{F}/\hat{\alpha}$, where

$$\alpha = \text{trace}[(W - W\hat{\Delta}(\hat{\Delta}' W\hat{\Delta})^{-1}\hat{\Delta}' W)\hat{\Gamma}]/r, \tag{17}$$

where r is the degrees of freedom of the model. The above scale-corrected chi-square statistic was introduced into covariance structure analysis by Satorra and Bentler (1988) (see also Satorra 1992 and Satorra and Bentler 1994). A mean- and variance-corrected chi-square may also be computed (e.g., see Satterthwaite 1941; Satorra and Bentler 1994).

To summarize this subsection on conventional estimation and testing, we distinguish three approaches to analysis. First, under normal theory analysis, parameter estimates are obtained using the NTML or NTGLS fitting functions, and standard errors and the chi-square test of model fit are computed using the conventional formulas of (14) and $n\hat{F}$. In ADF analysis, parameters are estimated using the weighted least squares fitting function of (6), setting W to the

ADF-type $\hat{\Gamma}$ of (11), standard errors are computed via (14), and a chi-square model test is obtained as $n\hat{F}$. With robust normal theory analysis, parameter estimates are obtained by the NTML or NTGLS estimators. Using the normal theory W and the ADF-type $\hat{\Gamma}$, standard errors are computed via (13), and the chi-square test of model fit is computed either as the residual chi-square of (15), or as the scaled chi-square of (17).

Muthén and Kaplan (1985, 1992) carried out Monte Carlo studies of normal theory analysis and ADF analysis using factor analysis on nonnormal data. They found that normal theory analysis gave good inferences for small models (around five variables), but inflated chi-square values and a downward bias in standard errors for larger models (ten or more variables). ADF analysis gave good standard errors and chi-square tests for small models and large samples (of at least size 1,000), while larger models did not show good results. Larger models produced inflated chi-square values and a downward bias in standard errors that was comparable to or worse than that of normal theory analysis. Apparently the asymptotic properties of ADF are not realized for the type of models and the finite sample sizes often used in practice. The method is also computationally demanding when there are many variables. This means that while ADF analysis may be theoretically optimal, it is not a practical method. Robust normal theory analysis appears to be an attractive alternative. To date, however, there is very limited experience with this approach. Muthén and Kaplan (1985, 1992) showed that normal theory estimates usually show very little bias even under nonnormality. A few studies with small models have recently reported promising results with regard to the robust standard errors and robust chi-square for nonnormal data (e.g., see Satorra and Bentler 1994; Chou, Bentler, and Satorra 1989; Satorra 1990), but nothing has been reported for models of realistic size. Robust normal theory analysis appears to warrant further study, and this will be carried out here in connection with complex sampling.

4. COVARIANCE STRUCTURE ANALYSIS FOR COMPLEX SAMPLES: AGGREGATED APPROACH

Consider now a sample of observations on y obtained under complex sampling. In this case iid observations cannot be assumed. As a

typical example, consider the data structure of the NHIS88, which in an idealized version can be described as follows: there are observations on a vector y_{ijkl}, where $i = 1, 2, \ldots I$ refers to strata; $j = 1, 2, \ldots, J$ refers to PSUs; $k = 1, 2, \ldots, K$ refers to segments; and $l = 1, 2, \ldots, L$ refers to households (with a single person observed per household). In this section, aggregated modeling will be considered; this is followed by disaggregated modeling in *Section 5*.

4.1. *Aggregated Modeling*

Aggregated structural equation modeling analysis refers to analysis of the conventional $\Sigma = \Sigma(\kappa)$ covariance structure model of (3). The inference procedures related to nonnormality that were discussed in the previous section also provide a useful basis for complex sample analysis. In particular, we propose that robust normal theory analysis can be generalized to complex samples. A consistent estimator of the Γ matrix can be formulated for complex sampling situations, and the use of this Γ estimator provides standard errors and chi-square tests of model fit that are robust not only to complex samples but also to nonnormality of the variables. We will first consider the estimation of Γ under a relatively simple complex sampling design and then introduce further complexities. The basic idea for estimating Γ is a classic and simple one that may be summarized as follows. When estimating a mean say, from complex sampling the mean estimate can be obtained as a linear combination of cluster means. If the clusters are independently sampled, the variance of the cluster means can be calculated using SRS formulas and applied to the sample mean. The structure of the sample below the cluster level need not be known.

Due to unequal selection probabilities and clustering, the Γ estimator of (11) is no longer consistent under complex sampling. Surveys provide weights corresponding to the inverse of the probability that a person or household is sampled. This probability is calculated as the product of conditional probabilities of selection at each stage of sampling. Usually, nonresponse, first-stage ratio adjustments, and poststratification ratio adjustments are also factored into the weights. In national samples, such inflation weights are used in Horvitz-Thompson type estimators such that the sum of the weighted sample observations provides an unbiased estimator of the population total. Equal weighting of the observations would bias the esti-

mate of the total because of such factors as oversampling of certain subpopulations.

In line with (9), we define instead a weighted $p^* \times 1$ data vector d_{ij} for stratum i and PSU j, using weights w_{ijkl} corresponding to the inverse selection probabilities,

$$d_{ij} = \sum_{k=1}^{K} \sum_{l=1}^{L} W_{ijkl} \begin{pmatrix} (y_{ijkl1} - \bar{y}_1)(y_{ijkl1} - \bar{y}_1) \\ (y_{ijkl2} - \bar{y}_2)(y_{ijkl1} - \bar{y}_1) \\ (y_{ijkl2} - \bar{y}_2)(y_{ijkl2} - \bar{y}_2) \\ \vdots \\ (y_{ijklp} - \bar{y}_p)(y_{ijklp} - \bar{y}_p) \end{pmatrix} \tag{18}$$

where the $\bar{y}$ variables refer to the elements of the p-dimensional weighted mean vector,

$$\bar{y} = n^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} w_{ijkl} y_{ijkl}, \tag{19}$$

and n is taken to be the sum of the weights instead of the total number of observations. The reduced vector of sample covariance matrix elements for the whole sample, taking weights into account, is

$$s_T = n^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J} d_{ij}, \tag{20}$$

where s_T denotes the $p^* \times 1$ vector of distinct elements of the weighted (total) sample covariance matrix S. For stratum i, consider the mean vector,

$$d_i = J^{-1} \sum_{j=1}^{J} d_{ij}. \tag{21}$$

Assuming that samples selected within different strata are independent and that the d_{ij} variables are iid within strata,

$$\text{var}(s_T) = n^{-2} \sum_{i=1}^{I} J\text{var}(d_{ij}), \tag{22}$$

where an unbiased and consistent estimator of $\text{var}(d_{ij})$ is obtained as

$$\hat{\text{var}}(d_{ij}) = (J - 1)^{-1} \sum_{j=1}^{J} (d_{ij} - d_{i.})(d_{ij} - d_{i.})'. \tag{23}$$

The variance estimator defined by (22) and (23) is a special case of the "non-parametric" variance estimator of Skinner et al. (1989, pp. 46–47) and the "random group estimator" of Wolter (1985, p. 33). It was recently proposed for covariance structure modeling in complex samples by Satorra (1990). The variance estimator gives an ADF-type $\hat{\Gamma}$. This means that robust, or distribution-free, standard errors and chi-square tests of model fit can be obtained with normal theory estimates in the way described in the previous section, providing a robust normal theory analysis suitable for complex samples. The great advantage of this estimator is that details of the sampling within PSUs need not be taken into account, because of the aggregation to the PSU level in (18). In terms of the introductory literature overview, this approach to estimating standard errors and computing chi-square may be characterized as a Taylor linearization method. The approach is general in that it can be used for parameters of any model that fits into the structural equation modeling framework.

Consider now some complications in the use of this estimator of Γ. It is clear from (23) that at least two PSUs per stratum need to be available. If some strata have only one PSU, as is the case with "self-representing" strata, such strata may be combined or split into random parts. Alternatively, second-level cluster variables, such as NHIS segments, may be redefined as PSUs (Parsons 1990).

The assumption that the d_{ij} variables are iid within strata is reasonable if PSUs are selected with replacement within each stratum with constant probability (WR sampling). This assumption is violated, however, when PSUs are selected as in the NHIS, using PSU selection without replacement with unequal probabilities proportional to size (UNEQWOR sampling). Parsons, Chan, and Curtin (1990), using NHIS data, compared the estimates of standard errors for totals and means obtained using WR and UNEQWOR procedures. They found that the WR approach overestimated the standard errors for means and proportions by up to 20 percent in some cases, although the bias was typically less than 10 percent. For totals the overestimation was considerably more severe. The issue of how to properly take ratio-adjustments in poststratification into account was discussed in Parsons and Casady (1986) and Parsons et al. (1990). Their results suggest that ignoring the effects of poststratification on variance estimation may not be serious.

Consider the estimation of Γ with sampling of PSUs within strata without replacement and with unequal probabilities propor-

tional to size (UNEQWOR). Since the UNEQWOR feature is involved in the NHIS, as well as many other surveys, this case deserves special attention. Let π_{ij} denote the probability of selecting PSU j in stratum i and let $\pi_{ijj'}$ denote the joint probability of selecting PSUs j and j'. Assume that m_{ij} second-stage units are selected without replacement from each PSU, with M_{ij} population units available. Using notation similar to that of (18), define the data vector d_{ijkl}, where the last two indices refer to the second stage units,

$$d_{ijkl} = w_{ijkl} \begin{pmatrix} (y_{ijkl1} - \bar{y}_1)(y_{ijkl1} - \bar{y}_1) \\ (y_{ijkl2} - \bar{y}_2)(y_{ijkl1} - \bar{y}_1) \\ (y_{ijkl2} - \bar{y}_2)(y_{ijkl2} - \bar{y}_2) \\ \vdots \\ (y_{ijklp} - \bar{y}_p)(y_{ijklp} - \bar{y}_p) \end{pmatrix} \tag{24}$$

and consider the variance of

$$s_T = n^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} d_{ijkl} \tag{25}$$

Drawing on standard sampling theory using the Yates-Grundy estimator (e.g., see Cochran, 1977, p. 301; Wolter 1985, p. 15; Shah, et al. 1989, p. A-7), we obtain the variance estimator in the multivariate case,

$$\begin{aligned} \hat{\text{var}}(s_T) = n^{-2} \sum_{i=1}^{I} [& \sum_{j=1}^{J} \sum_{j'<j}^{J} w_{ijj'}(d_{ij} - d_{ij'})(d_{ij} - d_{ij'})' \\ & + \sum_{j=1}^{J} \pi_{ij} m_{ij} (1 - \frac{m_{ij}}{M_{ij}}) S_{ij}], \end{aligned} \tag{26}$$

where

$$w_{ijj'} = \frac{\pi_{ij}\pi_{ij}' - \pi_{ijj}'}{\pi_{ijj}'}, \tag{27}$$

d_{ij} is defined as in (18),

$$S_{ij} = (m_{ij} - 1)^{-1} \sum_{j=1}^{m_{ij}} (d_{ijkl} - d_{ij.})(d_{ijkl} - d_{ij.})' \tag{28}$$

and

$$d_{ij.} = (m_{ij})^{-1} \sum_{k=1}^{K} \sum_{l=1}^{L} d_{ijkl}. \tag{29}$$

This UNEQWOR variance formula may be further refined to better take into account the features of a particular application. For example, the NHIS involves both self-representing strata (in which a single PSU is selected with certainty) and non–self-representing strata (in which two PSUs are selected UNEQWOR). It also has three substrata within PSUs, as indicated in the introduction. For this situation Parsons and Casady (1986), and Massey et al. (1989) suggested a variance estimator similar to (26) with S_{ij} instead defined as a within substratum variance estimator. Non–self-representing strata contribute terms as in (26), while self-representing strata contribute only within-substratum variance terms (see Massey et al. 1989, pp. 31–32).

4.2. *Correlation Structure Analysis with Dichotomous Variables*

This section presents a counterpart to robust normal theory analysis that extends robust standard errors and chi-square tests of model fit to analysis of statistics other than the sample covariance matrix S for continuous variables (this discussion draws on Muthén 1992; see also Muthén 1993). For example, Muthén (1989*b*) proposed the use of tetrachoric correlations in the factor analysis of binary items. In Muthén (1978, 1984) a weighted least squares procedure was proposed for the estimation and calculation of standard errors and chi-square test of model fit. An estimated matrix Γ was computed as a consistent estimate of the asymptotic covariance matrix of the sample tetrachoric correlations (see Muthén 1978). In this way, the Muthén (1978) estimator is analogous to the ADF estimator for continuous variables. In practice it suffers from the same type of computational and statistical limitations for large models as ADF does. The approach presented here avoids these limitations.

In the dichotomous case, correlations are analyzed and there is no issue of scale dependency. Because of this, a simple analogue to robust normal theory analysis for continuous variables would be to obtain model parameter estimates by unweighted least squares, using $W = I$, the identity matrix, in (13), (14), and (17). Alternatively, for correlations with widely differing variability, we may use a W with standard deviations of the estimated correlations as diagonal elements and zeros elsewhere. It remains to determine a proper estimator of Γ.

Assuming underlying multivariate normality for a set of di-

chotomous variables y, Muthén (1978) considered the $[p + p(p - 1)/2]$—dimensional sample vector t of z values (thresholds) for each variable and tetrachoric correlations for each pair of variables created from the vector of univariate and bivariate proportions u. Asymptotically, the covariance matrix of t can be expressed as a function of the covariance matrix of u, as follows,

$$\text{var}(t) = [\frac{\partial \pi}{\partial \rho'}]^{-1}\text{var}(u)[\frac{\partial \pi}{\partial \rho'}]'^{-1}, \tag{30}$$

where ρ is the population vector of thresholds and tetrachoric correlations and π is the population vector of unvariate and bivariate probabilities. Muthén (1978) obtained an ADF-like weighted least squares analysis using a consistent estimator of var(t) under iid as $\hat{\Gamma}$.

Consider now the computation of $\hat{\Gamma}$ for analysis of tetrachoric correlations under complex sampling. This development will be limited to the case of WR sampling in line with (18)–(23). Using a p-dimensional observation vector y_{ijkl} of 0s and 1s, consider the $[p + p(p - 1)/2]$—dimensional data vector d_{ij},

$$d_{ij} = \sum_{k=1}^{K}\sum_{l=1}^{L} w_{ijkl} \begin{pmatrix} y_{ijkl1} \\ \vdots \\ y_{ijklp} \\ y_{ijkl2}y_{ijkl1} \\ y_{ijkl3}y_{ijkl1} \\ y_{ijkl3}y_{ijkl2} \\ \vdots \\ y_{ijklp}y_{ijklp-1} \end{pmatrix} \tag{31}$$

so that the vector of univariate and bivariate proportions may be expressed as

$$u = n^{-1}\sum_{i=1}^{I}\sum_{j=1}^{J} d_{ij} \tag{32}$$

Using the conventional estimator of var(d_{ij}) given in (23), the variance estimator for t is then obtained in line with (22) and (30) by inserting estimated parameters into the expression

$$\text{var}(t) = n^{-2}[\frac{\partial \pi}{\partial \rho'}]^{-1}(\sum_{i=1}^{I} J\,\text{var}(d_{ij}))[\frac{\partial \pi}{\partial \rho'}]'^{-1}. \tag{33}$$

This new approach to the analysis of tetrachoric correlations promises to provide a computationally efficient way of obtaining complex-sample-robust standard errors and chi-square tests of model fit. It is clear that the approach is directly generalizable to ordered categorical data in line with Muthén (1984).

5. DISAGGREGATED MODELING

As described in the literature review, disaggregated modeling approaches attempt to take the sample design into account in the model. Variance component models for clustered data are a typical example. Consider the same prototypical example as before, assuming observations on a p-dimensional vector y_{ijkl}, where $i = 1, 2, \ldots, I$ refers to strata; $j = 1, 2, \ldots, J$ refers to PSUs; $k = 1, 2, \ldots, K$ refers to segments; and $l = 1, 2, \ldots, L$ refers to households (with a single person observed per household). As a starting point, consider the simple mixed model

$$y_{ijkl} = \mu + a_i + b_{ij} + c_{ijk} + d_{ijkl}, \tag{34}$$

where μ is a vector of overall means; a_i is a vector of fixed, stratum-specific effects; and b, c, d are uncorrelated vectors of random effects corresponding to the PSU, segment and household levels of nesting, having zero means and covariance matrices to be estimated. The task of aggregated modeling may be viewed simply as that of obtaining proper estimates of μ, the a_is and their standard errors, taking account of differences across PSUs and segments, which are random given the cluster sampling. Disaggregated modeling, however, places special emphasis on estimating and comparing the components of variation at the various levels of the data hierarchy. The individual-level variation can be compared to the variation due to segments and PSUs. This individual-level variation may be viewed as disaggregated, purged of the sociodemographic differences that are involved in segment and PSU variation. In this way, the disaggregated variance component approach has a higher level of ambition than aggregated modeling.

The sizes of the PSU and segment variance components influence the sizes of the PSU and segment intraclass correlations (icc's; e.g., see Koch 1983; Skinner et al. 1989), taken as the ratio of between cluster variance to total variance. The intraclass correlation measures

the degree of similarity within the same cluster. The larger the intraclass correlation, the larger the deviation from the assumption of independence between observations and the larger the distortion of conventional iid-based inference procedures. This distortion is usually expressed as the design effect (deff). For a univariate mean estimator $\hat{\kappa}$ and single-stage cluster sampling, with clusters of equal size,

$$\text{deff} = \frac{\text{var}_C(\hat{\kappa})}{\text{var}_{SRS}(\hat{\kappa})} = 1 + (c - 1)\rho, \tag{35}$$

where var_C denotes the (correct) variance under cluster sampling, var_{SRS} denotes the variance assuming simple random sampling, c is the common cluster size, and ρ is the icc (e.g., see Cochran 1977, pp. 240–42; Skinner et al. 1989, ch. 2).

A similar formula was obtained by Scott and Holt (1982) for a linear regression slope with a single explanatory variable x. Using the Fuller and Battese (1973) regression model with a clustered residual structure, they found the following approximation to the design effect:

$$\frac{\text{var}_C}{\text{var}_{SRS}} = 1 + (c - 1)\rho_\epsilon \rho_x, \tag{36}$$

where ρ_ϵ and ρ_x are the iccs for the regression residual ϵ and the x variable, respectively. Comparing (35) and (36) gives an explanation for the commonly observed phenomenon that deffs are larger for means than for regression coefficients (e.g., see Skinner et al. 1989, p. 68).

Standard errors of conventional (SRS) analysis are underestimated as soon as positive iccs are observed. Cluster size is an important factor as well. For example, the NHIS has an average of six households per segment, so by (36) a segment icc for ϵ of 0.1 together with an icc for x of 0.2 would result in a deff of only 1.1. This means that the standard error from a conventional analysis underestimates the true value by only 5 percent. On the other hand, the PSU cluster size is on average in 240 the NHIS. With PSU icc values of the same magnitudes, this would give a deff of 5.8 and an underestimation of the standard error by about 58 percent. Design effects for multivariate analysis are given in Skinner et al. (1989, pp. 43–44), where it is noted that these have a direct influence on model testing

by the Wald statistic under iid. Much more, however, needs to be known about design effects in multivariate analyses.

Extensions of covariance structure modeling to clustered data have recently emerged. Multivariate variance component analysis analogous to (34), but with latent variable structures for the random effects, has been proposed by Goldstein and McDonald (1988), McDonald and Goldstein (1989), Lee (1990), Longford and Muthén (1990), Muthén and Satorra (1989), and Muthén (1990, 1991). These developments concern ML estimation under the assumption of multivariate normality. Since variables from different levels of clustering may be entered into the analysis, these approaches are also termed multilevel techniques. These approaches are highly relevant to the model-based analyses of disaggregated models envisioned in this chapter.

Consider the special case of factor analysis in (1), where $B = 0$ and assume for simplicity that the sampling design is single-stage cluster sampling. Let $g = 1, 2, \ldots, G$ denote the clusters and $i = 1, 2, \ldots, n_g$ denote the individual observations within clusters, where the n_g are the varying cluster sizes. Assume the multilevel factor model

$$y_{gi} = \nu + \Lambda_B \eta_g + \Lambda_W \eta_{gi} + \epsilon_g + \epsilon_{gi}, \tag{37}$$

$$\text{var}(y_{gi}) = \Sigma_T = \Sigma_B + \Sigma_W, \tag{38}$$

$$\Sigma_B = \Lambda_B \Psi_B \Lambda_B' + \Theta_B, \tag{39}$$

$$\Sigma_W = \Lambda_W \Psi_W \Lambda_W' + \Theta_W, \tag{40}$$

where the subscript B stands for across-cluster variation and the subscript W stands for within-cluster variation. Muthén (1991) studied an application of this model where g and i represented eighth-grade classrooms and students in U.S. schools, respectively, and the vector y contained mathematics achievement test scores. Math ability, as represented by η, can be divided into a class and student component. Between-class variation is largely due to tracking into classes based on previous math performance, and different curricular emphases following this tracking.

In this context it is interesting to contrast aggregated and disaggregated modeling. Note that if $\Lambda_B = \Lambda_W = \Lambda$, it follows that

$$\Sigma_T = \Lambda \Psi_T \Lambda' + \Theta_T, \tag{41}$$

where Ψ_T and Θ_T are sums of the corresponding between- and within-parameter matrices. This shows that a factor model holds true also for Σ_T. It has the same number of factors and the same loading pattern as on the within (and between) level. The multilevel analysis estimates the factor and residual covariance matrices Ψ and Θ for both the between and within level. A conventional, aggregated, analysis of Σ_T estimates the sum of the between and within matrices. The conclusions drawn may differ substantially. For example, Muthén (1991) considered an example with equal Σ_W matrices within a set of mathematics classes that varied considerably in class means. Reliability of variables estimated from the factor model was considerably smaller using the disaggregated, multilevel model than the conventional, aggregated model.

It can be argued that the inference should concern the within parameters in Ψ_W and Θ_W and not the (total) parameters in Ψ_T and Θ_T. Studying the within parameters is a way of disentangling heterogeneous subpopulations (see also Muthén 1989*a*). For example, in terms of the Muthén (1991) analysis of math achievement, the regression coefficient (factor loading) for each observed achievement score regressed on the factor is taken to be the same in each of the classrooms, but the intercepts vary over classrooms. In this application, the classrooms vary greatly in terms of their factor values due to selection of students into eighth-grade math classes and the intercepts increase with rising factor values due to greater opportunities for learning in the more advanced classes. Given this, the within-class regressions have a flatter slope than the overall regression where classroom is ignored. The reliability estimates from the overall analysis are then inflated in the sense that they are not valid for any of the classrooms while the within-class estimates are valid in this sense. In addition, the between-class (co-) variation of math scores obtained in such a disaggregated analysis informs about the nature of the heterogeneity among the classes.

On the other hand, interest may instead focus on a model that is not conditional on design features. The particular mix of subpopulations seen in the total population might be of primary interest. In this case Ψ_T and Θ_T would seem to be the appropriate parameters for the inference and an aggregated model estimated via S_T, as in a conventional analysis, would be appropriate. Note, however, that models that hold true in subpopulations do not necessarily hold true

in the total population (cf. Muthén 1989*a*), even when they have the same factor pattern. In terms of the math achievement example, it can be argued that the reliability estimates are not intended for assessing reliability within a given classroom but in an overall sense for the total population of students.

Following Muthén (1990), ML estimation of the disaggregated model under normality may be briefly described as follows. Assume independent observations $g = 1, 2, \ldots, G$ on the data vector,

$$d_g' = (y_{g1}', y_{g2}', \ldots, y_{gn_g}'), \tag{42}$$

where each y_{gi} is of length p. Assuming equality of Σ_W across groups, the mean vector and covariance matrix of d_g are

$$\mu_{d_g} = 1_{n_g} \otimes \mu_y, \tag{43}$$

$$\Sigma_{d_g} = [I_{n_g} \otimes \Sigma_W + 1_{n_g} 1_{n_g}' \otimes \Sigma_B], \tag{44}$$

where $\otimes$ denotes the (right) Kronecker product, I denotes the identity matrix, and 1 denotes a vector of unit elements. It may be shown (Muthén 1990) that the log likelihood can be expressed as

$$\sum_{g=1}^{G}\{\log \mid \Sigma_B + n_g^{-1}\Sigma_W \mid + \text{trace}[(\Sigma_B + n_g^{-1}\Sigma_W)^{-1}(\bar{y}_g - \mu)(\bar{y}_g - \mu)']\} + (n - G)\log \mid \Sigma_W \mid + (n - G)\text{trace}[\Sigma_W^{-1}S_{PW}], \tag{45}$$

where S_{PW} is the regular pooled-within sample covariance matrix,

$$S_{PW} = (n - G)^{-1}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{gi} - \bar{y}_g)(y_{gi} - \bar{y}_g)'. \tag{46}$$

Note that the sample statistics involved in (45) include not only S_{PW} but also the cluster means $\bar{y}_g$. The computations are considerably more involved than in conventional covariance structure modeling because there are G additional terms (see the first line of [45]).

With an unrestricted mean vector μ and equal cluster sizes n_1, $n_2, \ldots, n_G = c$ (balanced data), a greatly simplified expression is obtained, as (45) can be written as

$$G\{\log \mid \Sigma_W + c\Sigma_B \mid + \text{trace}[(\Sigma_W + c\Sigma_B)^{-1}S_B] - \log \mid S_B \mid - p\} + (n - G)\{\log \mid \Sigma_W \mid + \text{trace}[\Sigma_W^{-1}] - \log \mid S_{PW} \mid - p\}, \tag{47}$$

where S_B is the between covariance matrix,

$$S_B = (G - 1)^{-1}\sum_{g=1}^{G} n_g(\bar{y}_g - \bar{y})(\bar{y}_g - \bar{y})', \tag{48}$$

with $\bar{y}$ denoting the overall sample mean vector.

It has been proposed by Muthén (1990, 1991) that the fitting function (47) be used in the general unbalanced case, terming this estimator MUML (Muthén's ML-based estimator) to contrast it with the full information ML (FIML) estimator obtained via (45). In the unbalanced case, c of (47) corresponds to an average-like cluster size

$$c = [n^2 - \sum_{g=1}^{G} n_g^2][n(G-1)]^{-1} \tag{49}$$

Muthén (1989*a*, 1990) pointed out that MUML estimation can be carried out as a special case of ML analysis by conventional, multiple-group structural equation software. MUML is obviously identical to FIML for balanced data, in which case the particular multiple-group analysis by conventional software gives the correct standard errors and likelihood ratio chi-square test of model fit. MUML is still a consistent estimator with unbalanced data, but like several estimators discussed under design-based analysis, MUML is then a limited information moment-fitting estimator. Preliminary experience with real and simulated data indicates that MUML estimation performs very well in the unbalanced case and gives estimates close to those of FIML (e.g., see Muthén 1994*a*). In this way, MUML provides a computationally feasible estimator of multilevel covariance structure models. The estimator can be generalized to multiple stages of clustering in a straightforward fashion.

5.1. *Disaggregated Analysis Under Nonnormality*

As mentioned in the introductory literature review, model-based, disaggregated analysis is sensitive to model misspecification. One important example of possible misspecification is when the normal density assumed in (45) is used with nonnormal data. Here, we will discuss briefly a very simple, but potentially quite useful, approach to disaggregated analysis under nonnormality.

Current research (e.g., see Muthén 1991) suggests that conventional analysis of S_{PW} in (46) for single-stage cluster sampling gives estimates of within parameters that are very close to those provided by FIML for the multilevel model. This is a reasonable expectation since S_{PW} is an unbiased and consistent estimator of Σ_W (e.g., see Muthén 1990). Such an analysis provides estimates of the disaggrega-

ted within part of the multilevel model, while between parameters are not estimated. Given an interest in disaggregated modeling, one could argue that the within parameters are the most important ones of the multilevel model.

Under normality and iid assumptions, the sample covariance matrix S_{PW} follows a Wishart distribution with $n - G$ degrees of freedom. Because of this, conventional normal theory analysis via (4) or (5) is appropriate. Consider now an analysis of S_{PW} that is robust against both complex sampling and nonnormality of variables. To be comparable to the multilevel model discussion above, a single-stage cluster sampling procedure will be considered, with G clusters of size $n_1, n_2, \ldots, n_G$. We will term this robust normal theory analysis of S_{PW}. The estimates are obtained using (4) or (5) on S_{PW}. The robust standard errors and chi-square are obtained as follows. Define the data vector,

$$d_g = \sum^{G} n_g \begin{pmatrix} (y_{gi1} - \bar{y}_{g1})(y_{gi1} - \bar{y}_{g1}) \\ (y_{gi2} - \bar{y}_{g2})(y_{gi1} - \bar{y}_{g1}) \\ (y_{gi2} - \bar{y}_{g2})(y_{gi2} - \bar{y}_{g2}) \\ \vdots \\ (y_{gip} - \bar{y}_{gp})(y_{gip} - \bar{y}_{gp}) \end{pmatrix} \tag{50}$$

so that the vector of distinct elements of S_{PW} in (46) may be expressed as

$$s_{PW} = (n - G)^{-1} \sum_{g=1}^{G} d_g. \tag{51}$$

If independent observations in the clusters can be assumed, it follows in line with (23) that the covariance matrix can be consistently estimated as

$$\hat{\text{var}}(s_{PW}) = (n - G)^{-2} G(G - 1)^{-1} \sum_{g=1}^{G} (d_g - d_.)(d_g - d_.)', \tag{52}$$

where $d_.$ is the mean of d_g over the G observations. Robust standard errors and chi-square are then simply obtained as before via (13) and (15)–(17). This type of analysis can also be generalized to stratified multistage cluster-sampling designs. According to substantive considerations, a decision can be made about which sampling level should

be chosen for the pooled-within calculations in (50). A similar approach to robust standard errors and chi-square testing may be attempted for the MUML estimator.

5.2. *Design-Robust Disaggregated Modeling*

The introductory literature review makes it clear that while model-based disaggregated modeling has much to offer in terms of efficient estimation, it also suffers from a lack of robustness to misspecification, due to such features as nonnormality and population heterogeneity, among others. In the previous section, the case of S_{PW} was used to illustrate a simple type of model-based complex sample analysis that was robust against nonnormality. It is also desirable, however, that any model-based analysis be robust against the complexities of the design. It is unlikely that exactly the same covariance structure model holds within different strata and clusters, as was assumed in the multilevel modeling described above. The full-population model is perhaps best characterized as a mixture model involving heterogeneous subpopulations. In this section, model-based analysis is combined with the design-based features of weighting and the use of design covariates to address this type of model misspecification.

Ignoring unequal probability sampling may lead to incorrect inferences for the full population model. Taking the simple example of S_{PW} analysis further, we may use design weights in the computation of d_g in (50) to avoid biases related to unequal selection probabilities.

Covariance structure modeling may also use design variables as covariates following ideas described by Pfeffermann and LaVange (1989) for regression models. They used information on homogeneous clusters of households called enumeration districts, obtaining data characterizing the enumeration districts from the U.S. census, on variables such as median family income, proportion of nonwhites, poverty level, and urbanicity. In many applications it is reasonable to assume that model parameters vary as functions of such sociodemographic variables. Muthén (1989*a*) considered related ideas in connecting conventional structural equation modeling to multilevel modeling (see also Skinner 1986). In the factor analysis framework, design variables can be included as regressors in a MIMIC structural equation model (see Muthén 1989a), where the factors behind a set of response

variables are regressed onto the design variables. This provides a disaggregated model where the factor model is assumed to hold conditionally on the regressors, while unconditionally a simple factor model may not hold. In the MIMIC approach, the means can vary as functions of the regressors. Certain direct effects from the regressors to the response variables can also be identified and estimated, and such effects allow for variable-specific differences in means across the levels of the regressors (design variables).

The multilevel model discussed earlier can also incorporate design variables. As an example, Muthén (1990) included class-level information on teaching activities in a model on student-level performance. Essentially, this is an empirical Bayes approach, where parameters vary randomly across cluster units (e.g., see Muthén and Satorra, 1989). Covariance structure modeling that allows for variation in level-related parameters such as means and intercepts is feasible by the FIML/MUML approach (Muthén 1990, 1991). More research, however, will be needed in order to handle more general parameter variation, such as that used for measurement slopes.

6. MONTE CARLO EVALUATION OF ESTIMATORS FOR COMPLEX SAMPLES

The performance of the complex-sample structural equation analyses proposed in the previous sections has not, with very few exceptions, been examined on real or simulated complex sample data. It is important to study the proposed methods under conditions comparable to those encountered with large-scale surveys to see if they are feasible for practical use. When this type of methods investigation has been carried out, it has relied on subsampling from real datasets. There have been relatively few Monte Carlo studies (e.g., see Rust 1985; Flyer, Rust, and Morganstein 1989).

We sought to obtain a precise picture of the behavior of these estimators by choosing a model that is correct and varying its parameter values to study different population conditions. A Monte Carlo study was conducted to study the methods under such controlled settings. Data were generated according to a model with variance components corresponding to cluster levels. This data simulation scheme also enables a precise assessment of the multilevel, model-based approach.

Three methods will be studied for both regression and factor analysis models. First, normal theory estimation via the ML fitting function of (4) will be carried out, with calculation of standard errors and, for the factor analysis model, the chi-square test of model fit via conventional normal theory. This is the normal theory analysis referred to earlier. For simplicity, this analysis approach will be called Method 1. Second, robust normal theory analysis will be carried out with ML estimation of parameters, and complex sample standard error and chi-square calculations via (13) and (17). This approach will be called Method 2. Finally, multilevel analysis will be carried out via FIML under normality assumptions. Since the data correspond to the balanced case, this means that the fitting function of (47) will be used with its normal theory standard errors and likelihood-ratio chi-square test of fit. This approach will be called Method 3. Note that Method 1 and Method 2 estimate the aggregated model that holds for Σ_T, while Method 3 estimates a disaggregated model with both between and within parameters. Method 1 ignores complex sampling, Method 2 represents a design-based approach to standard errors and chi-square, while Method 3 represents a model-based approach.

To mimic large-scale surveys, the Monte Carlo study needed to utilize larger models and larger and more complex datasets than are used in methods illustrations commonly seen in the structural equation modeling literature. To provide realistic values for the simulation parameters, the NLS, NHIS85, and NHIS88 datasets were investigated with respect to intraclass correlations (iccs). These iccs were estimated by methods described in Muthén (1991).

There are large variations in the size of intraclass correlations size across surveys due to what is being measured, how the clusters are formed, and the nature of the populations being studied. As an example, consider the school intraclass correlations for a set of attitudinal variables related to career interests, estimated from the base-year NLS data. For these high school seniors, the iccs ranged from 0.00 to 0.03. Similarly, Muthén (1994b) found iccs ranging from 0.02 to 0.06 for attitudes toward mathematics among U.S. seventh to tenth graders, in a survey where students were sampled within schools. In contrast to this, Muthén (1991) found class iccs of about 0.5 with mathematics achievement for U.S. eighth graders and school iccs of about 0.15.

Variations in the size of the iccs are similarly wide in the

NHIS. To cover a range of survey applications, several different levels of iccs will be used in the Monte Carlo study, with particular emphasis on small iccs in situations with large cluster sizes.

The Monte Carlo simulations were constructed using a simplified version of the NHIS sampling scheme. Within each of a number of strata, two PSUs were selected iid. Within each PSU, a number of units were sampled iid. Balanced data with equal numbers of units within PSUs were then simulated. Equal probability sampling was performed throughout, so that sampling weights were one. The key complexity of the data occurs because of clustering. Effects of various sizes of intraclass correlations and cluster sizes are the main concern. Nonnormality of variables was, however, introduced to make the data more realistic. The data were generated in line with the model of (34), simplified as

$$y_{ijk} = b_{ij} + c_{ijk}, \tag{53}$$

so that strata had equal means and there was no further nesting within PSUs. The study varied the sizes of intraclass correlations and clusters. The number of strata was also varied. Both normally and nonnormally distributed variables were considered. A regression model and a factor analysis model were studied.

6.1. *Data-Generation Models*

The data were generated according to regression and factor analysis models. In the regression case, a vector y of five observed variables was used. The first variable was taken as the dependent variable, while the remaining four were taken as regressors. In the factor analysis case, a vector y of ten observed variables was used. This is a large number of variables in covariance structure simulation contexts. In real-data situations, ten variables may be considered a small- or medium-sized model. For each of the two model types, y is generated as a sum of two multivariate normal vector components, each with mean vector zero and its own covariance matrix.

Let the term between (B) refer to the b_{ij} component and let within (W) refer to the c_{ijk} component. Different models are obtained by different choices of the covariance matrices Σ_B and Σ_W for the two component vectors. Given this data-generation scheme, the multilevel model given previously describes the data correctly and fully.

We may therefore discuss the two models in terms of the between and within parts of the multilevel model. The sum of the between covariance matrix Σ_B and the within covariance matrix Σ_W will be referred to as the total covariance matrix, Σ_T. The Σ_W and Σ_B matrices will be chosen so that the models hold for Σ_T, Σ_W, and Σ_B, although not always with the same parameter values. This means that the model structure is correct in all cases considered. In the over-identified factor model, the chi-square variables therefore refer to central chi-square distributions.

Regression Models. Denoting the dependent variable by y and the 4×1 vector of regressors by x, the regression model used to generate data may be written in multilevel terms as

$$y_{gi} = \alpha + \beta'(x_g^* + x_{gi}^*) + \epsilon_g^* + \epsilon_{gi}^*, \tag{54}$$

where the asterisked x and ϵ components are independent between and within components of the respective variable vector x and variable ϵ. Note that this variance component model can be analyzed by the multilevel covariance structure methods presented earlier. Estimation of α is then ignored. The intraclass correlation is set at 0.4 for all x variables and at values of 0.05, 0.10, and 0.20 for ϵ. The R^2 value is set at 0.4 for Σ_T. This is accomplished by assuming β coefficients of 0.323 for all x variables, between and within correlations of 0.2 for all pairs of x variables, and between variances of 0.6 and within variances of 0.4 for all x variables. For the ϵ intraclass correlation (icc) of 0.05, the between- and within-R^2 values are 0.84 and 0.30, respectively. For the ϵ icc of 0.10, the corresponding values are 0.73 and 0.31. For the ϵ icc of 0.20, the values are 0.57 and 0.33.

Factor Analysis Models. The factor model used to generate the data may be written in terms of the multilevel model in (37), with between and within matrices given in (39) and (40), using $\Lambda_B = \Lambda_W = \Lambda$ and diagonal Θ_B and Θ_W matrices. The total covariance matrix Σ_T is given as in (41). A two-factor model was chosen. The loading matrix Λ was taken to have a simple structure, with the first five variables loading only on the first factor and the last five variables only on the second factor. All loadings were taken to be one. The two factors are correlated. The correlation between the factors was set at 0.5 for between, within, and total. All ten observed variables were taken to

have the same intraclass correlation, which took on the values 0.05, 0.10, and 0.20. The within parameters were held constant across icc values. All within residual variances were set at 4, within factor variances at 2, and the within factor covariance at 1. The variable-factor correlations were therefore all 0.58. For the icc of 0.05, all between residual variances were 0.25, between factor variances 0.0658, and the between factor covariance 0.0329. The variable-factor correlations were all 0.42. For the icc of 0.10, all between residual variances were 0.5, between factor variances 0.167, and the between factor covariance 0.0835. The variable-factor correlations were all 0.50. For the icc of 0.20, all between residual variances were taken to be 1, between factor variances 0.5, and the between factor covariance 0.25. The variable-factor correlations were all 0.58.

For the factor analyses, the metric of the factors was set by fixing a factor loading for each factor. In the multilevel model this was carried out on both levels. For the conventional factor model of Methods 1 and 2 which consider Σ_T, the number of parameters is 21 and the number of degrees of freedom is 34. For the multilevel factor model of Method 3, the number of parameters and degrees of freedom is twice as large since there is a model for both Σ_B and Σ_W.

6.2. *Nonnormality*

Nonnormality was introduced only into the factor analysis model. All ten of the simulated observed variables in the factor analysis model had five categories scored 0, 1, 2, 3, and 4 with category percentages 3, 6, 13, 30, and 48. The resulting variable may be viewed as a very skewed Likert item of a sort commonly observed in real data. This strong nonnormality is common in many types of studies. By choosing the same cut points for all ten variables in the factor model, the factor model structure is preserved for the nonnormal data (cf. Muthén and Kaplan 1985), although the model will have different parameter values than in the normal case.

6.3. *Number of Strata and Cluster Sizes*

To approximate the size of the NHIS88 data, 100 strata with two PSUs per stratum were considered. With 240 units in each of the 200 PSU clusters, 48,000 observations were obtained in all. This huge

sample size would correspond to the whole NHIS88 dataset. Often, however, subsamples of the data are analyzed: for example, we may study NHIS panels comprised of a quarter of the overall sample, or 12,000 observations. We may interpret such subsamples as a design with a smaller cluster size within PSUs. Because the size of the PSU clusters will strongly affect the estimation, this factor was varied. For the regression model, PSU cluster sizes of 7, 15, 30, 60, 120, and 240 were used. Since the factor model is computationally more cumbersome, only sizes 7, 15, 30, and 60 were studied there.

The different cluster sizes may be taken to represent different survey situations. The size of 7 units may be viewed as representing the average number of households per segment in the NHIS88. Sizes 15 and 30 may be taken as typical class sizes in educational applications, and also represent the number of students sampled per school in the NLS. Sizes of 60 and above can be taken to represent NHIS88 PSU sizes. The total number of clusters (PSUs) is likely to be a factor in how well the complex sample variance estimation procedures perform. To represent other surveys with fewer strata, cases with 2 PSUs per each of 25 strata were also studied.

6.4. *Results of the Monto Carlo Study*

The presentation focuses on chi-square and standard errors, although results on parameter estimates will also be given for Method 3 (multilevel analysis under normality) since little is known about this method. In structural equation modeling, correct estimation of chi-square is as important as correct estimation of standard errors. Demands for precise standard error estimation in structural models are probably less stringent than in typical survey analysis of means and totals. In line with this, the present report will not consider a bias in the standard error of less than 10 percent as practically significant.

Results will be discussed for the regression model first, followed by the factor model. For the factor model, results for normal data will be presented first, followed by those for nonnormal data. Through most of the presentation, cases of 100 strata with 2 PSUs per stratum (200 PSUs in total) will be discussed, but at the end a few important cases with 25 strata and 2 PSUs (50 PSUs in total) will also be covered. With 100 strata the samples sizes for cluster sizes 7, 15,

30, 60, 120, 240 are 1,400, 3,000, 6,000, 12,000, 24,000, and 48,000, respectively. One thousand replications for each situation were used throughout the study.

Tables 1 through 8 present the results of the Monte Carlo study. In these tables, the percentage bias in the estimated standard error of a parameter estimate is defined as $100(a-b)/b$, where a is the mean of the estimated standard error over the 1,000 replications and b is the parameter estimate standard deviation over the 1,000 replications. For the normal theory parameter estimation of Methods 1 and 2, no results on bias in the estimation of parameters will be reported, since biases are in all cases negligible (see also Muthén and Kaplan 1985, 1992). It should be noted that the three methods analyze exactly the same data for each situation studied.

Regression Model. Table 1 shows the results for the standard error of the slope of the regression model with normal data. Method 1 ignores the complex sampling features and as a result underestimates the standard error. As expected, the downward bias increases for increasing intraclass correlations and increasing cluster size. Note that even for the small icc of 0.05 the bias can get large with cluster

TABLE 1
Monte Carlo Summary for Regression Model with Normal Data
(Standard Error Bias % for Slope)

Residual Intraclass Correlation	Cluster Size					
	7	15	30	60	120	240
Method 1: Normal Theory Ignoring Complex Sampling						
0.05	−5	−13	−21	−33	−45	−58
0.10	−10	−20	−32	−46	−58	−66
0.20	−17	−31	−45	−58	−69	−78
Method 2: Robust Normal Theory						
0.05	−0	−1	−2	−3	−1	−2
0.10	−0	−1	−2	−2	−1	−2
0.20	−1	−0	−2	−1	−0	−2
Method 3: Multilevel Analysis Under Normality						
0.05	4	0	−2	0	0	−1
0.10	3	0	−3	0	0	1
0.20	2	0	−3	2	−2	−1

sizes above 30. For the smallest icc and cluster size combination, the bias can be considered negligible.

Method 2 (robust normal theory) dramatically improves on the standard error performance, and in all cases the remaining slight negative bias is quite acceptable. Cluster size appears to have little influence. In this model, Method 3 (multilevel analysis under normality) estimates the same slope parameter as Methods 1 and 2 and its standard error results are therefore directly comparable. Method 3 imposes an equality constraint on the between and within slopes in line with the model used to generate the data. Method 3 also gives a quite satisfactory performance. Its performance appears slightly worse for the smallest cluster size of 7. Methods 2 and 3 appear to be unaffected by the icc value.

Factor Analysis Model. The remainder of the results are related to the factor model. Table 2 gives the results of chi-square testing with normal data using Method 1 and Method 2. Four pieces of information are provided: the mean over 1,000 replications, the variance over 1000 replications, the reject proportion at an α level of 5 percent, and the reject proportion at an α level of 1 percent. Since the model has 34 degrees of freedom, the expected mean and variance of chi-square are 34 and 68, respectively. A key feature is the rejection rate at the common α level of 5 percent. With 1,000 replications, the 95 percent prediction interval around the expected value of 5.0 is 3.6 to 6.4.

The Table 2 results for Method 1 show that the combination of the smallest icc with the smallest cluster size gives a distortion of the conventional chi-square small enough that we may neglect it. For larger values, however, the distortion is quite severe. For the small icc of 0.05, a severe distortion is obtained at cluster size 60, a size often exceeded in large-scale surveys. For the smallest cluster size of 7, an icc exceeding 0.1 is needed to give a severe distortion.

Consider next the Table 2 chi-square results for Method 2, where complex sampling is taken into account. In comparison to Method 1, a dramatic improvement has taken place. The scaled chi-square approach used with Method 2 appears to overcorrect slightly, but quite satisfactory results are obtained. This is a very encouraging result, particularly since the model is relatively large.

TABLE 2
Monte Carlo Summary for Factor Model with Normal Data
(Chi-Square Tests [34 d.f.] Method 1 and Method 2)

Intraclass Correlation		Cluster Size			
	7	15	30	60	
	Method 1: Normal Theory Ignoring Complex Sampling				
0.05	Chi-Square				
	Mean	35	36	38	41
	Var	68	72	80	96
	5%	5.6	7.6	10.6	20.4
	1%	1.4	1.6	2.8	7.7
0.10	Chi-Square				
	Mean	36	40	46	58
	Var	75	89	117	189
	5%	8.5	16.0	37.6	73.6
	1%	1.0	5.2	17.6	52.1
0.20	Chi-Square				
	Mean	42	52	73	114
	Var	100	152	302	734
	5%	23.5	57.7	93.1	99.9
	1%	8.6	35.0	83.1	99.4
	Method 2: Robust Normal Theory Analysis				
0.05	Chi-Square				
	Mean	33	33	33	32
	Var	62	61	61	59
	5%	3.6	3.5	3.2	2.8
	1%	0.8	0.8	0.4	0.3
0.10	Chi-Square				
	Mean	33	32	32	31
	Var	62	59	57	56
	5%	3.2	3.3	2.2	2.4
	1%	0.7	0.6	0.2	0.1
0.20	Chi-Square				
	Mean	32	32	31	31
	Var	60	56	55	55
	5%	2.8	2.6	1.7	2.0
	1%	0.5	0.4	0.3	0.4

Table 3 gives the corresponding standard error biases for Method 1 and Method 2. Overall, the Method 1 distortions appear less dramatic than those for chi-square. The largest cluster size, however, always yields an important distortion, as is also true for the

TABLE 3
Monte Carlo Summary for Factor Model with Normal Data
(Standard Error Bias % Method 1 and Method 2)

Intraclass Correlation	Parameter	Cluster Size 7	15	30	60
	Method 1: Normal Theory Ignoring Complex Sampling				
0.05	λ	1	−3	−1	−8
	θ	−4	−4	−9	−12
	ψ	1	−2	−2	−7
0.10	λ	−1	−8	−10	−22
	θ	−6	−9	−19	−26
	ψ	−1	−7	−10	−17
0.20	λ	−9	−22	−31	−46
	θ	−12	−22	−36	−47
	ψ	−9	−22	−31	−45
	Method 2: Robust Normal Theory Analysis				
0.05	λ	1	−2	2	−2
	θ	−3	−2	−5	−4
	ψ	1	−0	1	−1
0.10	λ	1	−3	2	−2
	θ	−3	−3	−7	−4
	ψ	1	−2	1	−1
0.20	λ	1	−3	0	−2
	θ	−3	−4	−8	−5
	ψ	0	−3	1	0

largest icc value. The biases for the icc of 0.05 are much smaller for the factor model than for the regression model, but the two are not directly comparable because the simulation for the regression case had a high intraclass correlation among the x variables.

Consider next the standard error performance for Method 2. As for the regression model, a slight negative bias remains. The size of this bias, however, is negligible in all cases. Taken together with the chi-square results, we conclude that Method 2 performs very well.

Tables 4 and 5 pertain to Method 3 and the use of the multilevel factor analysis model for the case of normal data. In this case, the set of parameters is different and direct comparisons with Methods 1 and 2 are not possible. Instead, Method 3 should be judged on its own merits. Since the factor analysis parameter estima-

TABLE 4
Monte Carlo Summary for Factor Model with Normal Data (Parameter Estimate Bias % [First column] and Standard Error Bias % [Second column]; Method 3: Multilevel Analysis Under Normality)

Intraclass Correlation	Parameter	Cluster Size							
		7		15		30		60	
0.05									
	λ_W							0	1
	θ_W							−0	−5
	ψ_W							0	−1
	λ_B							4	−13
	θ_B							−3	−17
	ψ_B							11	−18
0.10									
	λ_W					0	2	0	0
	θ_W					−0	−4	−0	−0
	ψ_W					0	2	0	−1
	λ_B					6	−16	2	−8
	θ_B					−2	−6	−2	−5
	ψ_B					4	−4	6	−1
0.20	λ_W	0	−1	0	−2	0	2	0	1
	θ_W	−0	−1	−0	0	−0	−4	−0	−4
	ψ_W	0	−1	−0	2	−0	2	0	0
	λ_B	4	−11	3	−7	2	−5	0	−2
	θ_B	−2	−4	−1	−5	−1	−5	−1	−4
	ψ_B	4	−5	4	−7	2	1	3	3

tion in Method 3 is a quite recent development, results for parameter bias will be presented. To our knowledge, Method 3 factor analysis has not been studied at all in Monte Carlo studies and has had very little practical use to date. Convergence problems were frequently encountered in the simulations; for small icc values and small cluster sizes, we often obtained inadmissible negative estimates of the between variance components. Although reparameterizations and restricted estimation are in principle possible, this was not carried out here. Instead, Tables 4 and 5 report only on cases with larger icc values and larger cluster sizes.

Table 4 presents the Method 3 parameter and standard error biases for both within and between parameters in the case of normal data. The between parameters are the only ones that show parame-

TABLE 5
Monte Carlo Summary for Factor Model with Normal Data (Method 3: Multilevel Analysis Under Normality Chi-Square Tests [68 d.f.])

Intraclass Correlation		Cluster Size			
		7	15	30	60
0.05					
	Chi-Square				
	Mean				69
	Var				139
	5%				6.6
	1%				0.9
0.10					
	Chi-Square				
	Mean			69	69
	Var			146	140
	5%			5.8	6.3
	1%			1.6	0.9
0.20					
	Chi-Square				
	Mean	69	69	69	69
	Var	136	131	145	140
	5%	5.9	5.4	6.0	6.4
	1%	1.1	1.1	1.8	1.2

ter bias. This is to be expected, since they draw on the scarcer information available on between-cluster variation. The zero parameter bias of the within parameter estimates is in line with the lack of large-sample bias in the estimates from Methods 1 and 2. The bias in the between parameters appears to be consistently positive for λ and ψ, while negative for θ. On the whole, however, the Method 3 parameter estimate bias is negligible.

Table 4 also gives the Method 3 standard error results. Although the within parameters show negligible bias, the between parameters have a nonnegligible negative bias for smaller icc values. For the icc value of 0.20, however, the results are quite acceptable.

Table 5 gives the Method 3 chi-square results. Here, chi-square refers to the likehood-ratio chi-square under normality. Note that the number of degrees of freedom for the multilevel model is 68. It is seen that the Method 3 chi-square behavior is excellent.

Tables 6 and 7 report on the case of nonnormal variables. For

TABLE 6
Monte Carlo Summary for Factor Model with Nonnormal Data (Intraclass Correlation 0.10. Chi-Square Tests*)

	Cluster Size			
Chi-Square	7	15	30	60
Method 1: Normal Theory Ignoring Complex Sampling				
Mean	40	43	47	55
Var	94	101	125	182
5%	17.6	25.4	39.7	64.8
1%	6.7	9.9	19.4	42.8
Method 2: Robust Normal Theory Analysis				
Mean	33	32	32	31
Var	63	60	59	59
5%	3.3	2.7	2.2	1.7
1%	0.6	0.5	0.3	0.6
Method 3: Multilevel Analysis Under Normality				
Mean			73	75
Var			165	171
5%			12.6	14.3
1%			3.7	5.0

*The degrees of freedom are 34 for methods 1 and 2 and 68 for method 3.

TABLE 7
Monte Carlo Summary for Factor Model Nonnormal Data
(Intraclass Correlation 0.10. Standard Error Bias %)

	Cluster Size			
Parameter	7	15	30	60
Method 1: Normal Theory Under Complex Sampling				
λ	−12	−15	−19	−28
θ	−21	−26	−35	−42
ψ	−12	−15	−21	−29
Method 2: Robust Normal Theory Analysis				
λ	−1	−1	−0	−2
θ	−1	0	−3	3
ψ	1	0	−1	1
Method 3: Multilevel Analysis Under Normality				
λ_W			−12	−19
θ_W			−29	−36
ψ_W			−13	−20
λ_B			−26	−10
θ_B			−10	−9
ψ_B			−10	−6

simplicity, only the icc value of 0.10 will be used. To be precise, this icc value refers to the variables before categorization and will be somewhat different for the categorized variables. For example, in the case with the largest cluster size, the icc estimate for the categorized variables was 0.08.

Table 6 gives the chi-square results for all three methods. As expected, Method 1 gives strongly inflated values. By contrast with Table 2, the inflation is still strong for the smallest cluster size of 7, reflecting the added effect of nonnormality. Method 2 works very well in the nonnormal case. It is of great practical significance that this way of taking the complex sample into account can also protect against distortions due to deviations from normality. Method 3 is not expected to work well under nonnormality since it builds on the normality assumption in deriving the likelihood-ratio chi-square test. It appears even more sensitive to deviations from normality than Method 1.

Table 7 gives the standard error results for the nonnormal case, for all three methods. Methods 1 and 3 give severely biased results, while Method 2 again works well. Again, Method 3 was not expected to work well in this situation.

In Table 8 the number of clusters (PSUs) has been reduced from 200 to 50. The aim was to study how well Method 2 works when

TABLE 8
Monte Carlo Summary for Factor Model 50
PSUs with Normal Data (Method 2: Robust Normal Theory Analysis, Chi-Square Tests [34 d.f.]; and Standard Error Bias %)

	Cluster Size			
	7	15	30	60
Chi-Square				
Mean	34	33	33	32
Var	63	65	69	60
5%	4.7	3.8	3.6	2.3
1%	0.5	0.9	0.9	0.3
Standard Error Bias %				
λ	−4	−5	−0	2
θ	−4	−3	−7	−4
ψ	−7	−2	1	3

a smaller number of clusters is used to create the nonparametric variance estimator. In this case, normal variables are once again generated and an icc of 0.10 is used. The results are encouraging in that the behavior of both chi-square values and standard errors is still very good.

6.5. *Monte Carlo Conclusions*

The results of the Monte Carlo study are very enlightening. Both Methods 2 and 3 perform well. The poor performance of Method 1, which ignores the complex sample features, shows that complex sample methods like Methods 2 and 3 are needed. Much more research about Methods 2 (robust normal theory analysis) and 3 (multilevel analysis under normality) remains to be conducted, on topics such as design features (unequal probability sampling, weighting, and unbalanced data) and additional forms of statistical evaluation (such as coverage and power issues). The results so far are, however, very promising and provide a basis on which to develop further methods, including those discussed in the statistical methods section but not explored in the Monte Carlo study. Still more interesting statistical methods await development in the area of multivariate complex sample analysis.

Structural equation modeling using Method 2 promises to be a future standard for complex sample analysis and for the analysis of simple random samples. The simple and well-behaved normal theory estimates are used together with standard errors and chi-square computed via a nonparametric, or distribution-free variance estimator. This variance estimator is likely to perform well whenever there is a large enough number of clusters. This study suggests, however, that a number of clusters as small as 50 may be sufficient. Not only does this variance estimator protect against distortions due to complex sampling, but it also adjusts for deviations from normality in the variables.

Although Method 2 addresses the estimation of the usual covariance structure model for the whole population, Method 3 addresses a less aggregated model. Method 3 uses a multilevel model with parameters for different levels of the sampling. The distinction between within- and between-cluster parameters may be particularly interesting in cases of naturally occurring clusters, such as schools

and classes. The method may, however, be viewed simply as a way to disentangle population heterogeneity and focus the analysis on the disaggregated within-cluster parameters, while the between parameters merely describe the sampling procedure. If the model is correctly specified, this also provides a more efficient analysis than Method 2. The Monte Carlo study suggests that the Method 3 within parameters and their standard errors can be estimated very well for normally distributed variables. Much more research on statistical and computational matters is, however, warranted for these types of multilevel models.

7. SUMMARY

The statistical research of this study resulted in two potential solutions appropriate for analyzing complex sample data with covariance structure models. One is a general, aggregated, solution for handling complex sampling that is also robust against the nonnormal variable distributions that are often seen in survey data. The solution is generalizable to dichotomous and ordered categorical variables. Monte Carlo results suggest that this solution works well with both normal and nonnormal data, and also in situations with relatively few clusters and small cluster sizes. This solution can be used with many sampling designs and it improves analyses even under simple random sampling.

The other, disaggregated, solution takes a different approach to handling problems of complex sampling. It describes features of the cluster sampling using variance component parameters. Its strength is that it provides a more detailed description of the population. The Monte Carlo study results suggest that this solution behaves well with normal data. The solution is sensitive to violations of normality; however, this study suggests ways this solution might be made robust against nonnormality.

REFERENCES

Battese, G. E., R. M. Harter, and W. A. Fuller. 1988. "An Error-Components Model for Prediction of County Crop Areas Using Survey Satellite Data." *Journal of the American Statistical Association 83:* 23–36.

Bean, J. A. 1975. "Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples, an Empirical Comparison." Vital and

Health Statistics, Series 2, No. 65. U.S. Department of Health, Education, and Welfare. Washington: Government Printing Office.

Bentler, P. M. 1989. *EQS Structural Equations Program Manual.* Los Angeles: BMDP Statistical Software.

Binder, D. 1983. "On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review, 51* 279–92.

Bock, R. D. 1989. *Multilevel analysis of educational data.* San Diego: Academic Press.

Browne, M. W. 1982. "Covariance Structures." In *Topics in Applied Multivariate Analysis,* edited by D. M. Hawkins Cambridge: Harvard University Press.

———. 1984. "Asymptotically Distribution-Free Methods for the Analysis of Covariance Structures." *British Journal of Mathematical and Statistical Psychology* 37:62–83.

Bryk, A. S., and S. W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, Calif.: Sage Publications.

Chamberlain, G. 1982. "Multivariate Regression Models for Panel Data." *Journal of Econometrics* 18:5–46.

Chou, C. P., P. M. Bentler, and A. Satorra. 1989. "Scaled Test Statistics and Robust Standard Errors for Nonnormal Data in Covariance Structure Analysis: A Monte Carlo Study." Technical report, University of California, Los Angeles.

Cochran, W. G. 1977. *Sampling Techniques,* 3rd ed. New York: Wiley.

Diggle, P. J., K. Y. Liang, and S. L. Zeger. 1994. *Analysis of Longitudinal Data.* Oxford, England: Clarendon Press.

Durbin, J. 1967. "Design of Multistage Surveys for the Estimation of Sampling Errors." *Applied Statistics* 16: 152–64.

Ferguson, T. S. 1958. "A Method of Generating Best Asymptotically Normal Estimates with Application to the Estimation of Bacterial Densities." *Annals of Mathematical Statistics* 29:1046–62.

Flyer, P., K. Rust, and D. Morganstein. 1989. "Complex Survey Variance Estimation and Contingency Table Analysis Using Replication." Paper presented at the American Statistical Association Meeting, Washington.

Freeman, D. H. Jr., J. L. Freeman, D. B. Brock, and G. G. Koch. 1976. "Strategies in the Multivariate Analysis of Data from Complex Surveys II: An Application to the United States National Health Interview Survey." *International Statistical Review* 44 (3):317–30.

Fuller, W. A. 1975. "Regression Analysis for Sample Survey." *Sankhya C* 37:117–32.

———. 1987. *Measurement Error Models.* New York: Wiley.

Fuller, W. A., and G. E. Battese. 1973. "Transformations for Estimation of Linear Models with Nested-Error Structure." *Journal of the American Statistical Association* 68:626–32.

Fuller, W. A., W. Kennedy, D. Schnell, G. Sullivan, and H. J. Park. 1986. "PC CARP." Ames: Statistical Laboratory, Iowa State University.

Fuller, W. A., D. Schnell, G. Sullivan, and W. J. Kennedy. 1987. "Survey Variance Computations on the Personal Computer." Paper delivered at the 46th Session of the International Statistical Institute, Tokyo.

Goldstein, H. I., and R. P. McDonald. 1988. "A General Model for the Analysis of Multilevel Data." *Psychometrika* 53:455–67.

Hansen, M. H., W. G. Madow, and B. J. Tepping. 1983. "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys." *Journal of the American Statistical Association* 78(384):776–807.

Hidiroglou, M. A. 1974. "Estimation of Regression Parameters for Finite Populations." Ph.D. diss., Iowa State University.

Holt, D., T. M. F. Smith, and P. D. Winter. 1980. "Regression Analysis of Data from Complex Surveys." *Journal of the Royal Statistical Society* A(143):474–87.

Jöreskog, K. G., and D. Sörbom. 1989. *LISREL: A Guide to the Program and Applications,* 2nd ed. Chicago: SPSS.

Kish, L. 1965. *Survey Sampling.* New York: Wiley.

Kish, L., and M. R. Frankel. 1974. "Inference from Complex Samples." *Journal of the Royal Statistical Society* B(36):1–37.

Koch, G. G. 1983. "Intraclass Correlation Coefficient." *Encyclopedia of Statistical Sciences* 4:212–17.

Kovar, M. G. 1985. "Approaches for the Analysis of Data." In *National Center for Health Statistics Plan and Operation of the Hispanic Health and Nutrition Examination Survey, 1982–84, Vital and health statistics.* Series 1, No. 19. DHHS Pub. No. (PHS) 85-1321. Public Health Service. Washington. Government Printing Office.

Laird, N. M., and J. H. Ware. 1982. "Random-Effects Models for Longitudinal Data." *Biometrics* 65:581–90.

Landis, J. R., J. M. Lepkowski, C. S. Davis, and M. E. Miller. 1987. "Cumulative Logit Models for Weighted Data from Complex Sample Surveys." Paper presented at the American Statistical Association Meeting, San Francisco.

Lee, S. Y. 1990. "Multilevel Analysis of Structural Equation Models." *Biometrika* 77:763–72.

Lemeshow, S., and A. M. Stoddard. 1984. "A Comparison of Alternative Variance Estimation Strategies for Estimating the Slope of a Linear Regression in Sample Surveys." *Communications in Statistics-Simulation Computation* 13(2):153–68.

Little, R. J. A. 1983. "Estimating a Finite Population Mean from Unequal Probability Samples." *Journal of the American Statistical Society* 78 (383):596–603.

———. 1989. "Survey Inference with Weights for Differential Sample Selection or Nonresponse." Paper presented at the meeting of the American Statistical Association, Survey Research Methods Section, Washington.

Longford, N. T. 1987. "A Fast-Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Effects." *Biometrika* 74:817–27.

———. 1989. Standard Errors for the Means of Proficiencies in NAEP: Jackknife Versus Variance Components." Technical Report for the Educational Testing Service.

———. 1993. *Random Coefficient Models.* Oxford, England: Clarendon Press.

Longford, N. T., and B. Muthén. 1990. "Factor Analysis for Clustered Observations." *Psychometrika* 57:581–97.

Magnus, J., and H. Neudecker 1988. *Matrix Differential Calculus.* New York: Wiley.

Malec, D., and J. Sedransk. 1985. "Bayesian Inference for Finite Population Parameters in Multistage Cluster Sampling." *Journal of the American Statistical Association* 80:897–40.

Massey, J. T., T. F. Moore, V. L. Parsons, and W. Tadros. 1989. "Design and Estimation for the National Health Interview Survey, 1985–94." National Center for Health Statistics. Vital Health Statistics, 2 (110).

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models,* 2nd ed. London: Chapman and Hall.

McDonald, R. P. 1980. "A Simple Comprehensive Model for the Analysis of Covariance Structures: Some Remarks on Applications." *British Journal of Mathematical and Statistical Psychology* 33:161–83.

McDonald, R. P., and H. Goldstein. 1989. "Balanced Versus Unbalanced Designs for Linear Structural Relations in Two-Level Data." *British Journal of Mathematical and Statistical Psychology* 42:215–32.

Muthén, B. 1978. "Contributions to Factor Analysis of Dichotomous Variables." *Psychometrika* 43:551–60.

———. 1983. "Latent Variable Structural Equation Modeling with Categorical Data." *Journal of Econometrics* 22:48–65.

———. 1984. "A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators." *Psychometrika* 49:115–32.

———. 1987. *LISCOMP. Analysis of Linear Structural Equations with a Comprehensive Measurement Model. Theoretical Integration and User's Guide.* Mooresville, Ind.: Scientific Software.

———. 1989*a*. Latent Variable Modeling in Heterogeneous Populations. Presidential Address to the Psychometric Society, July, 1989." *Psychometrika* 54:557–58.

———. 1989*b*. "Dichotomous Factor Analysis of Symptom Data." *Sociological Methods and Research* 18:19–65.

———. 1989*c*. "Tobit Factor Analysis." *British Journal of Mathematical and Statistical Psychology* 42:241–50.

———. 1990. "Mean and Covariance Structure Analysis of Hierarchical Data." UCLA Statistics Series #62.

———. 1991. "Multilevel Factor Analysis of Class and Student Achievement Components." *Journal of Educational Measurement* 28 (winter):338–54.

———. 1992. "A New Inference Technique for Factor Analyzing Binary Items Using Tetrachoric Correlations." Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

———. 1993. "Goodness of Fit with Categorical and Other Nonnormal Variables." In *Testing Structural Equation Models,* edited by K. A. Bollen and J. S. Long 205–43. Newbury Park, Calif.: Sage.

———. 1994*a*. "Multilevel Covariance Structure Analysis." *Sociological Methods and Research* 22:376–98.

———. 1994*b*. "Latent Variable Modeling of Longitudinal and Multilevel Data." Presented at the Showcase Session, Section on Methodology, American Sociological Association, August, 1994.

Muthén, B., and D. Kaplan. 1985. A Comparison of Some Methodologies for the Factor Analysis of Nonnormal Likert Variables." *British Journal of Mathematical and Statistical Psychology,* 38, 171–89.

———. 1992. "A Comparison of Some Methodologies for the Factor Analysis of Non-normal Likert Variables: A Note on the Size of the Model." *British Journal of Mathematical and Statistical Psychology* 45:19–30.

Muthén, B., and A. Satorra. 1989. "Multilevel Aspects of Varying Parameters for the Factor Analysis of Nonnormal Likert Variables." In *Multilevel Analysis of Educational Data,* edited by R. D. Bock, 87–99. San Diego: Academic Press.

Nathan, G., and D. Holt. 1980. "The Effect of Survey Design on Regression Analysis." *Journal of the Royal Statistical Association* Series B, 42 (3):377–86.

Parsons, V. L. 1990. Personal communication.

Parsons, V. L., and R. J. Casady. 1986. "Variance Estimation and the Redesigned Health Interview Survey." *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 406–17.

Parsons, V. L., J. Chan, and L. R. Curtin. 1990. "Analytic Limitations to Current National Health Surveys." Draft, National Center for Health Statistics.

Pfeffermann, D., and D. J. Holmes. 1985 "Robustness Consideration in the Choice of a Method of Inference for Regression Analysis of Survey Data." *Journal of the Royal Statistical Association,* Series A, 148,(3):268–78.

Pfeffermann, D., and L. LaVange. 1989. "Regression Models for Stratified Multistage Cluster Samples." In *Analysis of Complex Surveys,* edited by C. J. Skinner, D. Holt, and T. M. F. Smith, 237–60. West Sussex, England: Wiley.

Prosser, R., J. Rasbash, and H. Goldstein. 1990. *ML3-Software for Three-Level Analysis: Users' Guide.* London: Institute of Education.

Rao, J. N. K., and D. R. Thomas. 1988. "The Analysis of Cross-classified Categorical Data from Complex Sample Surveys." In *Sociological Methodology 1988,* edited by Clifford C. Clogg, 213–69. Washington: American Sociological Association.

Rao, J. N. K., and C. F. J. Wu. 1985. "Inference from Stratified Samples: Second-Order Analysis of Three Methods for Nonlinear Statistics." *Journal of the American Statistical Association* 80(391):620–30.

———. 1988. "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83:231–41.

Raudenbush, S., and A. Bryk. 1988. "Methodological Advances in Studying Effects of Schools and Classrooms on Student Learning." *Review of Research in Education* 15:423–75.

Rust, K. 1985. "Variance Estimation for Complex Estimators in Sample Surveys." *Journal of Official Statistics* 1:381–97.

Rutter, C. M., and R. Elashoff. 1994. "Analysis of Longitudinal Data: Random Coefficient Regression Modeling." *Statistics in Medicine* 13:1211–31.

Satorra, A. 1989. "Alternative Test Criteria in Covariance Structure Analysis: A Unified Approach." *Psychometrika* 54:131–51.

———. 1990. "Robustness Issues in Structural Equation Modeling: A Review of Recent Developments." *Quality and Quantity* 24:367–86.

———. (1992). "Asymptotic Robust Inferences in the Analysis of Mean and Covariance Structures." In *Sociological Methodology 1992* edited P. Marsden, 249–78. Oxford, England: Blackwell Publishers.

Satorra, A., and P. M. Bentler. 1988. "Scaling Corrections for Chi-Square Test Statistics in Covariance Structure Analysis." *1988 Proceedings of the Business and Economic Statistics Section of the American Statistical Association,* 308–11.

———. 1990. "Model Conditions for Asymptotic Robustness in the Analysis of Linear Relations." *Computational Statistics and Data Analysis* 10:235–49.

———. 1994. "Corrections to Test Statistics and Standard Errors in Covariance Structure Analysis." In *Latent Variable Analysis in Developmental Research,* edited by A. von Eye and C. Clogg, 285–305. Newbury Park, Calif.: Sage Publications.

Satterthwaite, F. E. 1941. "Synthesis of Variance." *Psychometrika* 6:309–16.

Schnell, D., H. J. Park, and W. A. Fuller. 1987. *EV CARP.* Ames: Statistical Laboratory, Iowa State University.

Scott, A. J., and D. Holt. 1982. "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods." *Journal of the American Statistical Association* 77:848–54.

Scott, A. J., and T. M. F. Smith. 1969. "Estimation in multistage surveys." *Journal of the American Statistical Association* 64:830–840.

Shah, B. V., M. M. Holt, and R. E. Folsom. 1977. "Inference About Regression Models from Sample Survey Data." Paper presented at International Association of Survey Statisticians Third Annual Meeting, New Delhi, India, December 5–15.

Shah, B. V., L. M. LaVange, B. G. Barnwell, J. E. Killinger, and S. C. Wheeless. 1989. *SUDAAN: Procedures for Descriptive Statistics, User's Guide.* Research Triangle Park, N.C.: Research Triangle Institute.

Skinner, C. J. 1986. "Regression Estimation and Poststratification in Factor Analysis." *Psychometrika* 51:347–56.

Skinner, C. J., D. Holt, and T. M. F. Smith, eds. 1989. *Analysis of Complex Surveys.* West Sussex, England: Wiley.

White, H. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50:1–25.

Wilson, M. 1989. "An Evaluation of Woodruff's Technique for Variance Estimation in Educational Surveys." *Journal of Educational Statistics* 14:88–101.

Wolter, K. M. 1985. *Introduction to variance estimation.* New York: Springer-Verlag.

Woodruff, R. S. 1971. "A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of the American Statistical Association* 66: 411–14.