STRUCTURAL EQUATION MODELING, *12*(3), 411–434

Sampling Weights in Latent Variable Modeling

Tihomir Asparouhov
Muthén & Muthén

This article reviews several basic statistical tools needed for modeling data with sampling weights that are implemented in M*plus* Version 3. These tools are illustrated in simulation studies for several latent variable models including factor analysis with continuous and categorical indicators, latent class analysis, and growth models. The pseudomaximum likelihood estimation method is reviewed and illustrated with stratified cluster sampling. Additionally, the weighted least squares method for estimating structural equation models with categorical and continuous outcomes implemented in M*plus* extended to incorporate sampling weights is also illustrated. The performance of several chi-square tests under unequal probability sampling is evaluated. Simulation studies compare the methods used in several statistical packages such as M*plus,* HLM, SAS Proc Mixed, MLwiN, and the weighted sample statistics method used in other software packages.

Unequal probability of selection is an inevitable feature of complex sampling surveys. This can be the result of stratified sampling, cluster sampling, subpopulation oversampling, designed unequal probability sampling, and so on. If the unequal probability of selection is not incorporated in the analysis, a substantial bias in the parameter estimates may arise. This bias is commonly known as selection bias. If the probability of selection is known and incorporated in the analysis, the selection bias can be eliminated. An unbiased estimator for the mean under unequal probability sampling was first developed by Horvitz and Thompson (1952). Skinner (1989) developed the pseudomaximum likelihood (PML) method, which under unequal probability sampling can be used to estimate any statistical model including the latent variable models discussed here.

This article describes several basic statistical tools needed to deal with unequal probability of selection that are implemented in M*plus* version 3 (Muthén &

Requests for reprints should be sent to Tihomir Asparouhov via e-mail: tihomir@statmodel.com

Muthén, 1998–2004; see http://www.statmodel.com). The article shows how the PML method is implemented in M*plus* version 3 for structural equation models and general latent variable models. In addition, it shows how the M*plus* implementation of weighted least squares (WLS) estimation for structural equation models with mixed outcomes (Muthén, 1984) is adapted to incorporate weighting for unequal probability of selection. Another purpose of this article is to promote simulation methods as a means for evaluation of statistical tools that incorporate weighting.

As a first step, the concept of unequal probability of selection and description of the framework for the simulation studies is clarified. Single-level models where no clustering or grouping information is available about the data and all sample units are considered independent, albeit not selected with equal probability is considered. Simulation studies on a factor analysis model, a latent class model, and a factor analysis model with binary indicators are conducted. The PML method, currently implemented in M*plus,* is compared with the commonly accepted practice of computing the weighted mean and covariance as a first step, followed by an analysis assuming simple random sampling as a second step. This method is called the *weighted maximum likelihood* (WML) method. For continuous outcomes, this method is implemented in various statistical packages (e.g., LISREL 8.51). The case of stratified cluster sampling is also compared, where samples are independent between clusters but not within clusters and are obtained from different strata. This methodology is illustrated with a simulation study on a binary factor analysis model. Finally, the performance of the methods implemented in the following statistical software packages is compared: M*plus,* MLwiN, and HLM/SAS Proc Mixed for a linear growth model.

DEFINITIONS AND INTERPRETATION

As a first step, the basic concepts related to unequal probability sampling and a simulation study that evaluates the performance of various statistical tools are described.

Let the probability of selection be p and the corresponding weight variable be $w = 1/p$. Let us first clarify the meaning of these quantities. If the population is infinite, for every individual the probability of selection is zero and the weight infinite; that is, these quantities are not well defined. One way to avoid this problem is to assume that the population is not infinite but a finite large population so that $p > 0$ and as a whole the large finite population would be numerically equivalent to an infinite population. Indeed many simulation studies in the complex sampling literature are designed that way; for example, Kaplan and Ferguson (1999) and Pfeffermann, Skinner, Holmes, Goldstein, and Rasbash (1998). Alternatively we can assume that the population is infinite and that p only represents a relative fre-

quency of occurrence in the sample. Thus the ratio of p_i:p_j would represent the sample frequency of occurrence of individuals similar to individual i to the sample frequency of occurrence of individuals similar to individual j, given that individuals i and j are equally common in the infinite target population. If in the infinite population the frequency of occurrence of individuals similar to individual i relative to the frequency of occurrence of individuals similar to individual j is represented by Q_i:Q_j and in the sample population that ratio is q_i:q_j, we can set $p_i = q_i/Q_i$. Thus we have two different methods to implement a simulation study: one using a finite target population and the second using an infinite target population. The two approaches, however, are in fact numerically equivalent. Indeed in what follows, only the relative value of p matters for the PML method and the p values can be standardized to some sort of arbitrary scale that is intuitive to understand but irrelevant to the statistical analysis. In addition, the finite target population approach with any estimation method converges to the infinite target population approach as the finite target population increases. To avoid the complexity of generating two population sets, target and sample, the infinite population simulation method is adopted. This eliminates the need for constructing the finite target population. We assume an infinite target population described by a model.

Let Y represent all dependent variables, X all predictor variables, and I the inclusion indicator; that is, I is a random variable that is 1 if (Y, X) is part of the sample and 0 otherwise. Simulation studies are conducted to determine the effects of sampling with unequal probability of selection on model estimation and inference. In addition to the usual model specification, a model for I is specified and I is sampled from that model. Individual elements (Y, X, I) are included in the final sample only if $I = 1$, and the information that will be available during the analysis is (Y, X, w), where $w = 1/p$ and $p = P(I = 1)$ is the probability of selection for included cases. The selection model can be defined by $P(I = 1) = f(Y, X)$ where f is some appropriate function or even more generally $P(I = 1) = f(Y, X, A)$ can be a function of Y, X and some auxiliary variables A that are not part of the model we are estimating. The variables A can be correlated with X or Y in the whole population or in a subpopulation only. In practical applications the weights are computed implicitly in terms of Y and X, and they are computed according to the sampling design specifications, which in turn can be connected to Y and X in a unknown way. We sample the infinite population until a predetermined number of sample units are included in the sample.

The sample selection is called noninformative if I and Y are conditionally independent given X; that is, the probability of selection p is a function of X and any other variable independent of Y but not of Y itself or an auxiliary variable A correlated with Y. When the selection is noninformative the true distribution of X is misrepresented in the sample, however a correctly specified conditional model $[Y|X]$ is estimated correctly even if no weights are included in the analysis. In fact, the inclusion of the weights in the analysis may result in a loss of estimation efficiency and the inference

may be less powerful (see Chambers, Dorfman, & Sverchkov, 2003). At present, however, there are no easily available tools that can establish the noninformativeness of the selection, and therefore one should not assume that this is the case, unless the computation of the weights involves variables that are already included in the analysis as predictors. If the selection is informative, however, a substantial bias in the parameter estimates may arise if the weights are not incorporated in the analysis. The focus here is primarily on informative selection mechanisms.

In practical applications the weights are the cumulative result of a comparison between the target population structure and the sample population structure, stratification, poststratification, and sampling design considerations. The sampling mechanism and the weights' computation can be very complex in practical application. The goal here is not to emulate such complex schemes, and thereby complicate the simulations, but to demonstrate the fundamental principles underlying the analysis of data obtained with unequal selection probabilities.

PSEUDOMAXIMUM LIKELIHOOD (PML)

The PML estimates are obtained by maximizing the weighted log-likelihood

$$\log(L) = \sum_i w_i \log(L_i) \tag{1}$$

where the subscript i runs over all independent observations. The asymptotic covariance matrix of these estimates is obtained by the sandwich estimator

$$\left(\partial^2(\log(L))/\partial\theta\partial\theta'\right)^{-1}\left(\sum_i w_i^2(\partial(\log(L_i))/\partial\theta)(\partial(\log(L_i))/\partial\theta)'\right)\left(\partial^2(\log(L))/\partial\theta\partial\theta'\right)^{-1} \tag{2}$$

where $\partial/\partial\theta$ and $\partial^2/\partial\theta\partial\theta'$ represent the first and the second derivative and the sum is over all individuals in the sample. Skinner (1989) showed that regardless of the choice of model, the PML parameter estimates are consistent under any sampling scheme.

It is the goal of this study to clearly demonstrate the difference between the PML method and the WML method described as follows: The WML parameter estimates are obtained by maximizing Equation 1 as well, but their asymptotic covariance matrix is estimated by the weighted information matrix

$$\left(\partial^2(\log(L))/\partial\theta\partial\theta'\right)^{-1} = \left(\sum_i w_i\partial^2(\log(L_i))/\partial\theta\partial\theta\right)^{-1}$$

The WML method applied to models with normally distributed outcomes amounts to computing the weighted sample statistics and fitting these sample statistics with

the usual maximum likelihood (ML) fit function. Later it is shown that the WML standard errors and confidence intervals are generally too short and that the classic inference based on the log-likelihood chi-square difference testing rejects more frequently than the designated rejection level. It is also demonstrated that the robust chi-square tests implemented in M*plus,* using correction factors, remain valid.

FACTOR ANALYSIS EXAMPLE

First the effects of selection bias for a factor analysis model with normally distributed outcomes were studied. Using an informative selection mechanism, the selection bias of the parameter estimates that results from ignoring the weights was determined, the underestimation of standard errors for the WML method was determined, and the effect of selection on several chi-square tests was considered. A factor model with five continuous and normally distributed variables Y_1, …, Y_5 and one factor η was considered. The mean of Y_j was ν_j and the residual variance θ_j. The variance of η is ψ. The loading of Y_j was λ_j and for identification purposes $\lambda_1 = 1$ was set. The selection mechanism was defined by $P(I = 1) = 1/(1 + e^{-Y_1})$; that is, the selection mechanism depended only on Y_1. This may correspond to a sampling situation where a particular factor measurement is considered to be very reliable and is used to deliberately oversample subpopulations with higher factor values. The analysis was replicated 500 times with 1,000 observations each.

Table 1 presents the parameter estimate bias (average estimate – true value) produced by the selection mechanism for the unweighted ML (ignoring the weights) analysis. The results show substantial selection bias for all parameters in the model with the exception of several residual variance parameters. The PML estimator, on the other hand, eliminates the bias completely.

Table 2 presents the effect of sampling weights on the standard errors computation. Here the parameter estimates were unbiased. This table presents the coverage probability of the 95% confidence intervals; that is, the probability that the confidence interval limits cover the true parameter value. If the methodology is correct and the sample size sufficiently large, that probability will be approximately 95%. The WML estimator clearly underestimates standard errors and confidence intervals and as a consequence produces low coverage probability. The coverage probability dropped by about 10% for most parameters and about 30% for the factor variance ψ. *T* tests or multivariate Wald tests based on such results would generally reject more often than they should. The last two columns in Table 2 show the ratio of standard deviation of the parameter estimates in the simulation to the average standard errors. This ratio should converge to 1 as the sample size increases if the asymptotic estimator is correct; however that was not the case for the WML estimator. The PML estimator, on the other hand, performed very well in terms of coverage and standard deviation to average standard errors ratio.

TABLE 1
Parameter Estimates Bias in Factor Analysis

Parameter	*True Value*	*PML/WML Bias*	*ML Bias*
λ_2	1	0.00	0.17
λ_3	1	0.01	0.18
λ_4	1	0.00	0.17
λ_5	1	0.00	0.18
ν_1	0.3	0.00	0.60
ν_2	0.3	0.00	0.27
ν_3	0.3	0.00	0.26
ν_4	0.3	0.00	0.27
ν_5	0.3	0.00	0.27
ψ	0.8	0.01	–0.28
θ_1	1	–0.01	–0.15
θ_2	1	0.00	0.00
θ_3	1	–0.01	–0.01
θ_4	1	0.00	0.00
θ_5	1	0.00	0.00

Note. PML = Pseudomaximum Likelihood; WML = Weighted Maximum Likelihood; ML = Maximum Likelihood.

TABLE 2
Standard Error Coverage in Factor Analysis

Parameter	*Coverage PML*	*Coverage WML*	*Coverage ML*	*SD/SE PML*	*SD/SE WML*	*SD/SE ML*
λ_2	0.946	0.816	0.406	1.08	1.56	0.98
λ_3	0.928	0.816	0.348	1.11	1.60	1.04
λ_4	0.952	0.822	0.404	1.05	1.53	0.94
λ_5	0.912	0.774	0.394	1.14	1.64	1.03
ν_1	0.938	0.740	0.000	1.05	1.83	1.01
ν_2	0.968	0.874	0.000	0.96	1.30	0.97
ν_3	0.964	0.850	0.000	0.99	1.36	0.98
ν_4	0.944	0.844	0.000	1.02	1.39	1.02
ν_5	0.944	0.844	0.000	1.05	1.44	1.05
ψ	0.898	0.658	0.010	1.21	2.28	1.01
θ_1	0.896	0.776	0.112	1.17	1.75	0.99
θ_2	0.936	0.866	0.946	1.08	1.32	1.02
θ_3	0.912	0.858	0.946	1.03	1.25	1.03
θ_4	0.920	0.864	0.946	1.10	1.35	1.05
θ_5	0.934	0.886	0.946	1.03	1.25	0.96

Note. PML = Pseudomaximum Likelihood; WML = Weighted Maximum Likelihood; ML = Maximum Likelihood.

For this particular example the WML estimator was equivalent to the ad-hoc method of first computing the weighted sample mean, variance, and covariance and then estimating the model parameters by the ML method based on these sample values. Another interpretation of the WML method is that it incorrectly treats the selection probability weights as frequency weights.

Next the one-factor model covariance structure was tested against the unrestricted covariance structure. This test has 5 *df.* Tables 3 and 4 present the chi-square results obtained by four different estimators, all of which incorporate the selection weights. The estimators MLR (Robust Maximum Likelihood), MLM (Mean-Adjusted Maximum Likelihood), and MLMV (Mean- and Variance-Adjusted Maximum Likelihood) are implemented in M*plus* and provide robust chi-square tests. The WML estimator uses the usual log-likelihood difference chi-square test statistic. All four estimators use the PML parameter estimates. The MLR estimator uses the PML asymptotic covariance matrix defined previously and a test statistic that is asymptotically equivalent to the T_2^* test statistic of Yuan and Bentler (2000). The asymptotic covariance matrix of the MLM and MLMV estimators is described in Muthén and Satorra (1995). The MLM chi-square test is

TABLE 3
Chi-Square Rejection Rates at the 5% Level

Sample Size	*MLR*	*MLM*	*MLMV*	*WML*
200	0.132	0.072	0.058	0.228
500	0.090	0.072	0.052	0.258
1000	0.078	0.054	0.044	0.256
2000	0.060	0.048	0.042	0.314
5000	0.062	0.046	0.046	0.318

Note. MLR = Robust Maximum Likelihood; MLM = Mean-adjusted Maximum Likelihood; MLMV = Mean- and Variance-adjusted Maximum Likelihood; WML = Weighted Maximum Likelihood.

TABLE 4
Chi-Square Test Statistic Average Value

Sample Size	*MLR*	*MLM*	*WML*
200	6.248	5.521	7.861
500	6.036	5.433	8.487
1000	5.603	5.197	8.754
2000	5.526	5.289	9.306
5000	5.140	5.024	9.382

Note. MLR = Robust Maximum Likelihood; MLM = Mean-adjusted Maximum Likelihood; MLMV = Mean- and Variance-adjusted Maximum Likelihood; WML = Weighted Maximum Likelihood.

the Satorra and Bentler (1988) chi-square, which is asymptotically equivalent to the MLR chi-square. Both of these estimators have mean corrected chi-square tests. The MLMV test statistic is the Satterthwaite (1946) variance correction of the MLM test statistic. The correction factors for these chi-square tests depend on the asymptotic covariance matrix. In fact the effect of the weights on the chi-square statistics is only indirect. The weights affect the asymptotic covariance, which in turn affects the correction factors for the chi-square statistics.

The results clearly show that only the robust chi-square tests MLR, MLM, and MLMV are acceptable. The WML chi-square difference test overrejects about five-fold. All three robust estimators performed well for sample sizes of 1,000 and above. For smaller sample sizes the MLMV outperformed MLM, which in turn outperformed MLR. The average values of the test statistics presented in Table 4 should converge to 5 as the sample size increases, because there were 5 *df*. The MLMV average value was not directly comparable to the degrees of freedom because its degrees of freedom were being estimated and changed from one replication to another. The WML average value was off even asymptotically, whereas the MLM and the MLR average chi-square statistics converged to their expected value.

LATENT CLASS ANALYSIS EXAMPLE

A simulation on a latent class analysis (LCA) model with two latent classes $C = 1$ and $C = 2$, and five observed dichotomous indicators U_1, …, U_5 was performed next. The dichotomous indicators took values 0/1 and $P(U_j = 0|C = k) = 1/(1 + e^{-\tau_{jk}})$. A predictor X of the categorical latent variable, with a standard normal distribution, was added to the model. Thus $P(C = 1|X) = 1/(1 + e^{-(\alpha + \beta X)})$. The sample consisted of 1,000 independent observations and this analysis was replicated 500 times. Half of the observations were selected using simple random sampling (SRS) and the other half were selected using the following selection model: $P(I = 1) = 1/(1 + e^{-2.5+\Sigma U_j})$. The selection model was chosen in this way to clearly produce informative sampling. A similar model and estimation technique are discussed in Patterson, Dayton, and Graubard (2002).

Table 5 shows the parameter estimate bias produced by this selection mechanism when the unweighted ML method is used. Almost all parameter estimates had substantial bias; for example, given that $X = 0$, the probability that an individual belongs to Class 1, as reflected by the α parameter, was underestimated by 11%. As expected, the PML estimator that incorporates the weights in the analysis essentially removes the bias completely.

It is well known that the Pearson and the likelihood ratio chi-square test statistics are affected by the selection mechanism as well. The proper corrections are described in Rao and Thomas (1989), for example. Under SRS the expected value of both statistics was 20 in this model because there were 20 *df*. The average value, however, for both statistics in this simulation under the unequal selection mecha-

TABLE 5
Parameter Estimates Bias in Latent Class Analysis

Parameter	*True Value*	*PML/WML Bias*	*ML Bias*
τ_{11}	1	0.06	–0.18
τ_{21}	0.3	0.02	–0.22
τ_{31}	0.3	0.03	–0.21
τ_{41}	1	0.04	–0.22
τ_{51}	1	0.04	–0.23
τ_{12}	–1	0.02	–0.14
τ_{22}	–1	0.02	–0.15
τ_{32}	–1	0.01	–0.15
τ_{42}	0	0.01	–0.14
τ_{52}	0	0.01	–0.13
α	0	0.09	–0.43
β	1.2	–0.09	–0.10

Note. PML = Pseudomaximum Likelihood; WML = Weighted Maximum Likelihood; ML = Maximum Likelihood.

nism was close to 30. Thus the chi-square tests were biased upward and were likely to reject more often than the target level. Table 6 shows that the WML estimator again produced standard errors and confidence intervals that are too short. The percentage of time the true parameters were covered by the WML confidence intervals dropped much below the desired 95% level.

The parameter estimate bias results of Table 5 were produced by the informative selection mechanism. If the selection is noninformative the bias would not exist. However the results of Table 6 can be replicated even with a noninformative selec-

TABLE 6
Standard Error Coverage in Latent Class Analysis

Parameter	*Coverage PML*	*Coverage WML*	*SD/SE PML*	*SD/SE WML*
τ_{11}	0.950	0.828	1.00	1.55
τ_{21}	0.942	0.746	1.05	1.73
τ_{31}	0.954	0.778	0.99	1.62
τ_{41}	0.970	0.834	0.97	1.47
τ_{51}	0.976	0.848	0.94	1.42
τ_{12}	0.952	0.892	0.95	1.18
τ_{22}	0.948	0.924	0.90	1.04
τ_{32}	0.964	0.936	0.95	1.11
τ_{42}	0.946	0.938	0.95	1.02
τ_{52}	0.946	0.928	0.98	1.05
α	0.966	0.816	0.92	1.39
β	0.960	0.834	1.04	1.52

Note. PML = Pseudomaximum Likelihood; WML = Weighted Maximum Likelihood.

tion mechanism. Even if the selection is noninformative, the fact that SRS is not used, but an unequal selection probability sampling is used instead, would cause the WML information matrix to produce incorrect results. The Pearson/log-likelihood chi-square test would also be biased upward even with a noninformative selection mechanism. This point is illustrated in Table 7. The same model is used but the selection mechanism was defined by $P(I = 1) = 1/(1 + e^{-X})$. This selection is noninformative and as expected all three estimators—ML, WML, and PML—showed almost no selection bias at all. The coverage for the ML and the PML estimators was also very good, but the coverage for the WML estimator dropped by about 10% for some parameters and the standard errors were underestimated by about 40%. The most efficient estimator in the noninformative case as expected was the ML estimator, which had about 20% shorter confidence intervals than the PML estimators. The mean squared error (MSE) for the ML estimator was also about 30% smaller than the MSE for the PML estimator. The advantage seen for the ML estimator over the PML and the WML estimators shows that with noninformative selection, ignoring the weights is better than including them. The phenomenon described in this paragraph can be observed in the factor analysis model from the previous section if we add a covariate X to that model as well.

The unweighted analysis is consistent when the selection is noninformative. However, the unweighted estimation of a summary value (e.g., the true proportion of individuals belonging to Class 1), which requires averaging over all covariates, can be biased because the distribution of the covariates can be misrepresented in the unequally selected sample.

Quite often the data for LCA are available in the form of a frequency table. However, it is not possible to summarize weighted data in a frequency table. Al-

TABLE 7
Bias and Coverage in Latent Class Analysis With Noninformative Selection

Parameter	*Bias ML*	*Bias PML and WML*	*Coverage ML*	*Coverage PML*	*Coverage WML*
τ_{11}	0.02	–0.01	0.958	0.968	0.932
τ_{21}	0.00	–0.01	0.944	0.952	0.944
τ_{31}	0.00	–0.01	0.930	0.938	0.920
τ_{41}	0.01	0.00	0.956	0.950	0.952
τ_{51}	0.01	–0.01	0.952	0.966	0.952
τ_{12}	–0.03	–0.09	0.942	0.946	0.828
τ_{22}	–0.01	–0.05	0.958	0.960	0.844
τ_{32}	–0.01	–0.04	0.926	0.966	0.856
τ_{42}	0.00	–0.02	0.962	0.960	0.834
τ_{52}	0.00	–0.01	0.958	0.964	0.852
α	–0.01	0.07	0.960	0.968	0.880
β	0.03	0.04	0.942	0.944	0.876

Note. ML = Maximum Likelihood; PML = Pseudomaximum Likelihood; WML = Weighted Maximum Likelihood.

though the correct parameter estimates can be found from such a table, the correct standard errors cannot. If the frequency table is computed by adding the weights over the same multivariate outcomes, and the robust MLR method is applied, the standard error will again be underestimated. The reason is that the sum of the weights is not used in the computation of the standard errors but rather the sum of the squares of the weights is used. Therefore if weighted categorical data have to be summarized in a frequency table, two tables should be produced: one for the sum of the weights and the second for the sum of the squares of the weights.

WEIGHTED LEAST SQUARES (WLS)

This section shows how the WLS approach of Muthén (1984) and Muthén, du Toit, and Spisic (1997) for estimating structural equation models with categorical and continuous outcomes can be adapted to include unequal selection probabilities. The treatment here also applies to the other three M*plus* estimators—WLSM (Mean-adjusted Weighted Least Square), WLSMV (Mean- and Variance-adjusted Weighted Least Square), and ULS (Unweighted Least Square)—that use similar techniques.

Denote by σ_1 the first-stage parameters (intercepts, thresholds, and slopes) and by σ_2 the second-stage parameters (correlations). Let $l_{ij} = L(Y_j|X)$ and $l_{ijk} = L(Y_j, Y_k|X)$ be the univariate and the bivariate conditional log-likelihoods for the *i*th individual's outcomes Y_j and Y_k. The total univariate and bivariate conditional pseudo log-likelihoods are $l_j = \sum_{i=1}^{n} w_i l_{ij}$ and $l_{jk} = \sum_{i=1}^{n} w_i l_{ijk}$. The first-stage estimates σ_1 are obtained by maximizing l_j. These estimates are the univariate PML estimates and are therefore consistent according to Skinner (1989). The second-stage estimates σ_2 are obtained by maximizing l_{jk}, given that the univariate parameters are fixed to their first-stage estimates. We call these the quasi-PML estimates, as opposed to the true PML estimates that would be obtained by maximizing l_{jk} over both univariate and bivariate parameters simultaneously. As in Muthén and Satorra (1995), under the regularity conditions B1 through B7, the consistency of the first-stage estimates and the consistency of the second-stage PML estimates implies the consistency of the quasi-PML estimates σ_2. Finally the third-stage estimates (i.e., the structural equation parameter estimates) are obtained by minimizing the objective function

$$F(\theta) = \sum \left(\sigma(\theta) - \hat{\sigma}\right) W^{-1} \left(\sigma(\theta) - \hat{\sigma}\right)', \tag{3}$$

where **W** is the weight matrix. The four different estimators that are implemented in M*plus* via this method differ in their choice of **W**. The third-stage estimates are consistent by Theorem 4.1.1. in Amemiya (1985).

The proof of the asymptotic normality is the same as in Muthén and Satorra (1995). Let

$$g_i = \left(\frac{\partial l_{i1}}{\partial \sigma_{1,1}}, \cdots, \frac{\partial l_{ip}}{\partial \sigma_{1,p}}, \frac{\partial l_{i21}}{\partial \sigma_{2,21}} \cdots, \frac{\partial l_{ipp-1}}{\partial \sigma_{2,pp-1}} \right)$$

be the complete score vector, where p is the length of the observed vector Y. Let $g = \sum_{i=1}^{n} w_i g_i$ be the total score vector. Let $\hat{\sigma}$ be the first- and the second-stage estimates and let $\bar{\sigma}$ be the true parameter value. As in Muthén and Satorra (1995) for some point σ^* between $\hat{\sigma}$ and $\bar{\sigma}$

$$0 = g(\hat{\sigma}) = g(\bar{\sigma}) + (\hat{\sigma} - \bar{\sigma}) \cdot \partial g(\sigma^*) / \partial \sigma$$

and therefore

$$n^{1/2} (\hat{\sigma} - \bar{\sigma}) = \left(\frac{-n^{-1} \partial g(\sigma^*)}{\partial \sigma} \right)^{-1} n^{-1/2} g(\bar{\sigma})$$

By the central limit theorem $n^{-1/2} g(\bar{\sigma})$ is asymptotically normal with mean zero and variance approximated by

$$V = n^{-1} \sum_{i=1}^{n} w_i^2 g_i(\bar{\sigma}) g_i(\bar{\sigma})' \tag{4}$$

If

$$A = p \lim \left(\frac{-n^{-1} \partial g(\sigma^*)}{\partial \sigma} \right) = p \lim \left(\frac{-n^{-1} \partial g(\bar{\sigma})}{\partial \sigma} \right)$$

we get that $n^{1/2} (\hat{\sigma} - \bar{\sigma}) \xrightarrow{d} N(0, \Gamma)$ with $\Gamma = A^{-1} V A'^{-1}$.

The structural parameters θ are estimated in the third stage by minimizing the objective function (Equation 3). We apply Theorem 4.1.3. in Amemiya (1985) to obtain the asymptotic distribution of the structural parameters $\hat{\theta}$

$$\mathrm{Var}(\hat{\theta}) = n^{-1} (\Delta' W^{-1} \Delta)^{-1} \Delta' W^{-1} \Gamma W^{-1} (\Delta' W^{-1} \Delta)^{-1}$$

where $\Delta = \partial \sigma / \partial \theta$.

Again it is clear from these formulas that although the parameter estimates are the same regardless of whether the weights are used as frequency weights or as unequal selection weights, the standard errors are not. If the weights are unequal selection

weights and are used as frequency weights the standard error/confidence interval size would be underestimated and the chi-square statistic would be inflated.

BINARY FACTOR ANALYSIS EXAMPLE

Next the performance of the WLS estimation of a factor analysis model with one factor η and five binary indicators $U_1, \ldots, U_5$ was evaluated. The variance of η was fixed to 1 and the distribution of U_j was defined by $P(U_j = 0|\eta) = \Phi(\tau_j - \lambda_j\eta)$, where Φ was the standard normal distribution function. The selection mechanism was defined by $P(I = 1) = 1/(1 + e^{2r+1.5-\Sigma U_j})$ where r was a uniformly distributed random number in the interval [0, 1]. Because of the random effect r the weights in this example were not connected to the data in a deterministic way. This simulation again used a sample size of 1,000 and results were accumulated over 500 replications.

First the WLSMV estimator (see Muthén, 1998–2004), was compared, with and without the weights. Table 8 shows the parameter estimates bias when the weights were omitted and the coverage of the true parameters by the 95% confidence intervals. The results clearly show that substantial selection bias arises if the weights are not incorporated in the analysis. The low coverage for the unweighted analysis is mostly due to the parameter estimates bias. On the other hand the weighted analysis had virtually no bias and the coverage was on target as well.

Next the performance of the chi-square statistics of the four WLS-based estimators available in M*plus*—WLS, WLSM, WLSMV, and ULS—was compared (see Muthén, 1998–2004). For all estimators, weights were incorporated in the analysis. All four estimators performed very well. Table 9 shows that the rejection rates

TABLE 8
Bias and Coverage in Binary Factor Analysis With WLSMV
(Mean- and Variance-adjusted Weighted Least Square)

Parameter	*True Value*	*Bias Weighted*	*Bias Unweighted*	*Coverage (SD/SE) Weighted*	*Coverage (SD/SE) Unweighted*
λ_1	0.8	0.04	–0.17	0.964 (1.173)	0.634 (1.340)
λ_2	0.8	0.01	–0.20	0.954 (0.999)	0.392 (1.002)
λ_3	0.8	0.01	–0.20	0.960 (0.992)	0.390 (1.008)
λ_4	0.8	0.01	–0.18	0.940 (1.038)	0.486 (1.005)
λ_5	0.8	0.01	–0.18	0.952 (1.054)	0.500 (1.015)
τ_1	–1.0	–0.03	–0.62	0.950 (1.274)	0.002 (1.658)
τ_2	0.3	0.00	–0.62	0.944 (1.021)	0.000 (0.986)
τ_3	0.3	0.00	–0.62	0.964 (0.997)	0.000 (0.998)
τ_4	1.0	0.00	–0.60	0.946 (1.007)	0.000 (0.980)
τ_5	1.0	0.01	–0.60	0.942 (1.042)	0.000 (0.998)

TABLE 9
Chi-Square Rejection Rates at the 5% Level

Sample Size	*WLS*	*WLSM*	*WLSMV*	*ULS*
200	—	0.052	0.042	0.053
500	0.060	0.052	0.052	0.040
1000	0.066	0.064	0.064	0.042

Note. WLS = Weighted Least Square; WLSM = Mean-adjusted Weighted Least Square; WLSMV = Mean- and Variance-adjusted Weighted Least Square; ULS = Unweighted Least Square.

were on target for all estimators. The WLS estimator with a sample size of 200 did not complete all 500 replications because of singular weight matrix and that value is not reported.

STRATIFIED CLUSTER SAMPLING

Stratification is used to reduce the variance of estimators by dividing the population into more homogeneous strata. Clustering, although increasing the variance of estimators, is used for logistic reasons to make sampling practical. This section gives the variance of estimators under both stratification and clustering. If cluster sampling is not taken into account, the standard errors will be underestimated and if the stratified sampling is not taken into account the standard errors will be overestimated.

When the data are obtained by stratified cluster sampling the observations that belong to the same cluster may not be independent and observations obtained from different strata may follow different distributions. The aggregate modeling approach does not model this dependence and difference in the distribution but rather estimates the parameters assuming that the observations are independent. An extensive discussion on this approach is available in Muthén and Satorra (1995), section 4. Means, variance, covariances, and other general structural equations can be estimated by ML ignoring the dependence. This method is called the quasi ML (QML) method. For example the QML estimates are consistent for a linear growth model and the preceding binary factor model. There are multilevel models, however, that cannot be aggregated. Models with random slopes are such models because the residual variance would not be constant but a function of the covariates. Nevertheless, even in such models, most of the parameter estimates are actually consistent.

The consistency of the QML estimation can be combined with unequal probability of selection and still produce consistent estimates simply by maximizing the weighted QML, which is called the quasi-PML (QPML) method. Note that the QPML parameter estimates are the same as the PML parameter estimates assum-

ing independence of the observations. The consistency of the QPML method extends to the WLS estimators in M*plus*.

The variance estimator of the QPML parameter estimates was adjusted according to Skinner (1989) to reflect the dependence between the observations and the stratification. Thus the variance estimator (Equation 2) is replaced with

$$\left(\partial^2(\log(L))/\partial\theta\partial\theta'\right)^{-1}\left(\sum_h \frac{n_h}{n_h-1}\sum_c (z_{ch}-\bar{z}_h)(z_{ch}-\bar{z}_h)^T\right)\left(\partial^2(\log(L))/\partial\theta\partial\theta'\right)^{-1} \quad (5)$$

where n_h is the number of sampled clusters from stratum h, $z_{ch} = \Sigma_i w_{ich}$ $\partial(\log(L_{ich}))\partial\theta$ is the total score for all individuals i in cluster c in stratum h and $\bar{z}_h$ is the average of z_{ch}. Similarly the variance of the WLS estimates was adjusted. The variance estimator (4) was replaced by

$$V = n^{-1}\sum_h \frac{n_h}{n_h-1}\sum_c \left(v_{ch}-\bar{v}_h\right)\left(v_{ch}-\bar{v}_h\right)', \quad (6)$$

where $v_{ch} = \Sigma_i w_{ich} g_{ich}(\bar{\sigma})$ was the total score vector for all individuals i in cluster c in stratum h and $\bar{v}_h$ was the average of v_{ch}. The remaining part of the WLS variance estimation was unchanged. These variance corrections are implemented in M*plus* version 3.1. A more extensive discussion of the effect of stratified sampling on SEM is available in Asparouhov (2004a).

FACTOR ANALYSIS WITH CLUSTER SAMPLING

A simulation study with cluster sampling was conducted to evaluate the performance of the WLS method and the variance correction introduced in Equation 6. For this purpose the binary factor model was modified to include a cluster effect. The population in this example consisted of 1 stratum with 50 clusters each of size 20. Let the factor η be decomposed as a within and a between part $\eta = \eta_w + \eta_b$, where η_w is an individual-level factor and η_b is a cluster-level factor. For each observed variable U_j the underlying latent variable was denoted by $U_j^* = \lambda_j\eta + \varepsilon_j$ where ε_j was a standard normal random variable. Thus $U_j = 0$ if and only if $U_j^* < \tau_j$. The between–within decomposition can be extended to $U_j^* = U_{wj}^* + U_{bj}^*$ where $U_{wj}^* = \lambda_j\eta_w + \varepsilon_j$ and $U_{bj}^* = \lambda_j\eta_b$. The following selection model was considered. All clusters were sampled at random and within each cluster the observations were sampled according to the following selection model: $P(I=1) = 1/(1+e^{-U_{w1}^*})$. This selection mechanism was identical across clusters and depended only on the within component of the underlying latent variable of the first measurement. It was assumed that the $\mathrm{Var}(\eta_w) = \mathrm{Var}(\eta_b) = 0.5$, so that the estimated model assuming $\mathrm{Var}(\eta) = 1$ was still correct regardless of the fact

that the observations were not really independent. Two estimators were considered. The first estimator was the WLSMV estimator ignoring the cluster sampling and the second estimator was WLSMV-complex, which uses the variance correction introduced in Equation 6. The rest of the model parameters were unchanged. Note that the variance correction changes not only the variance of the parameter estimates but also the parameter estimates because the weight matrix **W** changes. In general that change is very small and it disappears asymptotically. Indeed, the parameter estimates converged to the true value regardless of what the weight matrix was.

Table 10 clearly shows that both WLSMV and WLSMV-complex eliminated the selection bias through the weighting. When cluster sampling is ignored, the standard error and confidence intervals are underestimated. The standard errors of the loading parameters λ were underestimated by about 20% and those of the thresholds τ by about 80%. On the other hand the variance correction accounted properly for the dependence between the observations and provided correct standard errors, confidence intervals, and coverage. Although in this particular example the chi-square statistic produced correct rejection rates even when the clustering information was ignored, in general that would not be the case. The performance of the chi-square test is closely connected in general to the performance of the standard errors. For example, if instead of testing this model against an unrestricted model it was tested against a more restricted model that holds all thresholds equal, there would again be higher rejection rates from the multivariate Wald test as well as from the chi-square test, which in general produce close results, because the variance of the estimates is underestimated under the independence assumption.

TABLE 10
Bias and Coverage in Binary Factor Analysis With Cluster Sampling

Parameter	*Bias WLSMV*	*Bias WLSMV Complex*	*Coverage WLSMV*	*Coverage WLSMV Complex*	*SD/SE WLSMV*	*SD/SE WLSMV Complex*
λ_1	0.02	0.03	0.916	0.952	1.19	1.05
λ_2	0.01	0.01	0.888	0.930	1.20	1.02
λ_3	0.00	0.00	0.900	0.950	1.21	1.04
λ_4	0.01	0.01	0.904	0.940	1.17	0.99
λ_5	0.00	0.00	0.890	0.930	1.25	1.06
τ_1	0.02	0.02	0.760	0.956	1.68	1.01
τ_2	0.00	0.00	0.684	0.942	1.98	1.04
τ_3	0.00	0.00	0.682	0.932	2.00	1.04
τ_4	0.01	0.01	0.794	0.948	1.57	1.01
τ_5	0.00	0.00	0.784	0.948	1.61	1.04

Note. WLSMV = Mean- and Variance-adjusted Weighted Least Square.

COMPARISON AMONG M*PLUS,* MLWIN, AND HLM/SAS PROC MIXED

A simulation study compared the estimation methods for weighted data analysis implemented in M*plus* version 3 published by Muthén and Muthén, HLM 5.04 published by Scientific Software International, and MLwiN 1.1 published by the Institute of Education, University of London. The difference among the three approaches has been of interest for some time on Internet discussion lists, and conducting this simulation study has become increasingly important due to many scientists reporting large differences in the final parameter estimates and even in the significance of effects. M*plus* implements the PML method of Skinner (1989) and MLwiN implements the PWIGLS method of Pfeffermann et al. (1998). We verified that SAS Proc Mixed Version 9 implements the same method as HLM 5.04 and conducted the simulation study only with HLM but the results apply to SAS Proc Mixed as well. A partial reference for this method is available in Raudenbush, Bryk, and Congdon (2002) and the SAS manual. The method amounts to multiplying sample statistics by the weights, just as what is done for linear regression models.

A model that can be estimated by all statistical packages was selected, namely a linear growth model with normally distributed outcomes. However any other two-level hierarchical linear model could be used in this comparison. The following unbalanced design was used: 500 univariate observations were clustered within 100 Level 2 units. Half of the Level 2 units had four observations and the other half had six observations. The times of the observations were equally spaced starting at 0 and ending with 3 for the clusters with four observations and ending with 5 for the clusters with six observations. The linear growth model has a random intercept I and a random slope S for the time covariate t. Thus the observed variable Y_{it} for Level 2 unit i at time t satisfies

$$Y_{it} = I_i + S_i t + \varepsilon_{it},$$

where ε_{it} is a zero mean normally distributed residual with variance θ.The normally distributed random effects I_i and S_i have means μ_1 and μ_2, variances σ_1 and σ_2, respectively, and covariance ρ. The selection model is defined by the first measurement in the growth model Y_{i0}, namely $P(I_i = 1) = 1/(1 + e^{-Y_{i0}})$; that is, Level 2 units with higher initial status were oversampled. Thus the weights were computed by $w_i = 1/P(I_i = 1) = 1 + e^{-Y_{i0}}$ and were applied only at Level 2. This analysis was replicated 500 times.

The datasets were generated and then analyzed with each of the three statistical packages. The main input files used in this simulation for the three programs are provided in the Appendix. All files needed to replicate this simulation study in each of the three packages are also available on the M*plus* Web site (www.statmodel.com). These include all datasets formatted in three different ways

as used by the three packages, M*plus* input file, a DOS batch program that runs HLM with multiple datasets, and the MLwiN macro.

Table 11 shows the bias in the parameter estimates (average estimate – true value) as well as the coverage rates for the 95% confidence intervals computed by the three programs. The HLM estimates for μ_1, μ_2, σ_1, and ρ have substantial bias. The bias of the HLM estimate for μ_1 is about 20 times the bias of the M*plus* and MLwiN estimate. The coverage of the HLM estimator is quite low for some of these parameters and due primarily to poor parameter estimation. This example clearly demonstrates the severe flaws of that method. The HLM parameter estimates are in fact closer to the unweighted ML parameter estimates than to the true values. After reviewing an earlier draft of this work the HLM authors confirmed the existence of these flaws and released HLM Version 6 which implements the Pfeffermann et al. (1998) method. This method yields results almost identical to the results obtained by M*plus* 3.

The bias produced by M*plus* and MLwiN is virtually identical and is approximately zero. In fact the two programs produce almost identical results even over a single replication. The ML method is identical to the IGLS method in general. In fact the MLwiN results for unweighted analysis can be reproduced exactly in M*plus* by using the ML estimator with the expected information matrix option. However the PML and the PWIGLS methods do not produce identical parameter estimates and standard errors. Whereas the difference in the parameter estimates between the two methods is very small, the difference in the standard errors is not. The PML method clearly outperformed the PWIGLS method in terms of coverage of the true values.

This example is by no means extreme and the differences observed for this model among the three methods resemble those found in real data applications. The coverage for the theta parameter for the linear regression weighting method, currently implemented in HLM, was not reported because that value is not readily available in the HLM output. The σ_1 parameter has a relatively low coverage in M*plus*. That problem, however, disappears as the sample size increases.

The flaws of the method implemented in HLM/SAS are also exposed by the following observation. If an observation with a weight of 2 is replaced by two

TABLE 11
Bias and Coverage in HLM, MlwiN, and M*plus* for a Growth Model

Parameter	*True Value*	*HLM Bias*	*MlwiN Bias*	*Mplus Bias*	*HLM Coverage*	*MlwiN Coverage*	*Mplus Coverage*
μ_1	0.5	0.347	0.017	0.017	0.184	0.782	0.908
μ_2	0.1	0.077	0.002	0.002	0.710	0.888	0.942
σ_1	1.0	–0.139	–0.024	–0.024	0.850	0.758	0.848
σ_2	0.2	0.000	–0.006	–0.006	0.900	0.848	0.902
ρ	0.3	–0.062	–0.005	–0.006	0.850	0.846	0.940
θ	1.0	0.010	–0.008	–0.008	—	0.878	0.910

observations identical to it with weights of 1 the parameter estimates in HLM change. This is counterintuitive and has no interpretation. Both PML developed in Skinner (1989) and PWIGLS developed in Pfeffermann et al. (1998) claim to possess this multiplicative property. The unbiased multistage unequal probability sampling estimators for the mean also possess the multiplicative property. Note, however, that for single-level models or two-level models that can be interpreted as multivariate single-level models, the only method that satisfies the multiplicative property and is independent of the scale of the weights is the PML method. Other methods could satisfy the multiplicative property in some sort of asymptotic sense. The linear regression weighting method implemented in HLM, however, does not.

Finally, performance of the robust chi-square tests in M*plus* was studied. In M*plus* the preceding model can be estimated as a single-level multivariate model or as a multilevel univariate model. Both analyses were based on the PML method and produced the same results. In the multivariate case the units with four observations were appended with two more observations that are recorded as missing data. In the multivariate case M*plus* computes the chi-square test statistic for testing the estimated model against the unrestricted mean and covariance model. Just as in Table 4 for the factor analysis model, we can compare the performance of the three different test statistics available in M*plus* for this growth model. The MLM and MLMV estimators are available only for models without missing data. In the preceding model the missing data mechanism is missing completely at random and therefore listwise deletion with MLM or MLMV estimators are expected to produce valid results; however, they use only half of the sample. The results in Table 12 confirm the results in Table 3, namely that MLMV outperformed MLM, which in turn outperformed MLR. Although it is not possible to generalize the MLM estimator to the large class of models in which the MLR estimator is available, it is easy to develop Satterthwaite (1946) variance correction for the MLR estimator, which would substantially improve its performance, just as the variance correction of MLMV improves MLM.

TABLE 12
Chi-Square Rejection Rates at the 5% Level in Mplus for a Growth Model

Sample Size	*MLM*	*MLMV*	*MLR*
100	0.216	0.110	0.254
200	0.178	0.056	0.228
500	0.140	0.064	0.142
1000	0.112	0.058	0.128
2000	0.092	0.044	0.128
5000	0.076	0.030	0.106

Note. MLM = Robust Maximum Likelihood; MLMV = Mean-adjusted Maximum Likelihood; MLR = Mean- and Variance-adjusted Maximum Likelihood.

CONCLUSION

This article clarified the meaning of the statistical concept of weighting for unequal selection probability and also showed how effectively simulations can be used to evaluate the performance of statistical tools. Although there are various ways to conduct simulation studies with unequal selection probability sampling, the method adopted here is by far the simplest to implement and interpret. It may be limited in some ways in terms of generality. However, it is sufficient to provide an outlook on the quality of statistical tools. It is clear from the simulations reported here that omitting the weights can produce severely biased estimates for any latent variable model. No parameter appears to be immune from selection bias.

All statistical tools within M*plus* that can be used to analyze data with weights performed very well and according to the underlying theory. A single exception to this is the Pearson/likelihood ratio chi-square for mixture models with categorical data, which does not provide robust corrections similar to the chi-square tests available for the structural equation models. Our simulations also show that the MLR chi-square test requires a larger sample size than both MLM and MLMV. All testing procedures based on *t* tests or multivariate Wald tests perform as expected because they rely on the covariance estimation that accounts for the weighting.

The parameter estimates obtained by the PWIGLS method as implemented in MLwiN were very close to the PML estimates obtained in M*plus.* However, PWIGLS underestimates the asymptotic covariance matrix of the parameter estimates. It is not clear whether this is a flaw of the method in general or of the particular MLwiN implementation because the simulations reported in Pfeffermann et al. (1998) did not show this problem. The method implemented in the HLM and SAS Proc Mixed software did not perform well and produced substantial bias and low confidence interval coverage.[1] It was also clearly demonstrated that simply weighting the log-likelihood is not enough. It is not enough to simply compute the weighted sample statistics and analyze them assuming SRS. Note that this is exactly how weights are used with common structural equation software packages. A commonly accepted practice has been to compute the weighted sample mean and covariance and analyze them in a separate step assuming SRS (e.g., LISREL 8.51[2]). This study clearly demonstrated that this method overestimates the chi-square value and underestimates the asymptotic covariance of the parameter estimates, resulting in low confidence interval coverage. The problems of this method were also demonstrated in Kaplan and Ferguson (1999). In a somewhat different context Stapleton (2002) demonstrated this as well. The degree to which the WML method underestimates the

[1]The newly released HLM Version 6 implements the Pfeffermann et al. (1998) method. The results obtained by HLM 6 are almost identical to the results obtained by M*plus* 3. This suggests that the coverage problem of the MLwiN 1.1 implementation of the Pfeffermann et al. (1998) method may not be due to the method itself.

[2]The newly released LISREL Version 8.7 implements a maximum-likelihood based method that for most parameters closely agrees with the PML method implemented in M*plus* 3.

asymptotic covariance varies from one parameter to another and therefore it is not possible to mend this method by a simple adjustment of the sample size as in Potthoff, Woodbury, and Manton (1992). The same is true for categorical data analysis. It is not enough to compute the polychoric and polyserial correlations in one step and analyze their structure assuming SRS in a second step. Robust variance estimation and robust chi-square are necessary when weighted data analysis is performed and the weighting has to be accounted for.

An essential part of the analysis of unequal selection probability samples is the computation of the weights. This computation is as important as the rest of the analysis. Incorrect computation of the weights clearly can result in incorrect conclusions. The burden is on the researcher to understand the meaning of the weights and to guarantee their truthfulness. Although the computation of the weights does not involve any advanced mathematical algorithms, it can be increasingly complex as the sampling scheme becomes increasingly complex. At present the common statistical packages do not offer any help in computing the appropriate weights, but this part of the computation is by no means of lesser importance. Several practical complex sampling examples are described in Korn and Graubard (1999, Appendix A), where references are given to the detailed description of the weights computation. Such examples can guide applied researchers through this complex matter.

This article focused on the analysis of single-level multivariate models, which can also be interpreted as univariate two-level models. Asparouhov (2004b) offered a complete discussion on two-level and multilevel models, but we dismiss here any assumptions that the methods and results are somehow similar to what was described for the single-level analysis. The situation for two-level analysis is far more complicated and unsettled. The two methods currently available in statistical software for multilevel models—PML (M*plus*) and PWIGLS (MLwiN)—are based on the assumption that the sample size within each cluster converges to infinity, and that is a severe restriction, which has not been made clearly enough in the literature. The bias in the parameter estimates is unavoidable by both methods and it depends on how large the cluster sizes are and how informative the selection is. Consistent estimation for two-level models, with arbitrary cluster sizes, is simply not available at present.

ACKNOWLEDGMENT

I am thankful to Bengt Muthén for his guidance, to Linda Muthén for her support and commitment, and to Thuy Nguyen for computational assistance. I am thankful to Rod Little, Chris Skinner, Laura Stapleton, and the reviewers for helpful comments on an earlier draft of this article.

This research was supported by SBIR grant R43 AA014564–01 from NIAAA to Muthén & Muthén.

REFERENCES

Amemiya, T. (1985). *Advanced econometrics.* Cambridge, MA: Harvard University Press.

Asparouhov, T. (2004a). *Stratification in multivariate modeling* (M*plus* Web Note No. 9). Retrieved December 16, 2004, from www.statmodel.com

Asparouhov, T. (2004b). *Weighting for unequal probability of selection in multilevel modeling.* Manuscript submitted for publication.

Chambers, R., Dorfman, A., & Sverchkov, M. (2003). Nonparametric regression with complex survey data in analysis of survey data. In R. L. Chambers & C. J. Skinner (Ed.), *Analysis of survey data* (chapter 11). New York: Wiley.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47,* 663–685.

Kaplan, D., & Ferguson, A. J. (1999). On the utilization of sample weights in latent variable models. *Structural Equation Modeling, 6,* 305–321.

Korn, E. L., and Graubard, B. I. (1999). *Analysis of health surveys*. New York: Wiley.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

Muthén, B. O. (1998–2004). *Mplus technical appendices.* Los Angeles: Muthén & Muthén.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Accepted by *Psychometrika.*

Muthén, B., & Satorra, A. (1995). Technical aspects of Muthén's Liscomp approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika, 60,* 489–503.

Muthén, L. K., & Muthén, B. O. (1998–2004). *Mplus user's guide* (3rd ed.). Los Angeles: Muthén & Muthén.

Patterson, B. H., Dayton, C. M., & Graubard, B. I. (2002). Latent class analysis of complex sample survey data: Application to dietary data. *Journal of the American Statistical Society, 97,* 721–734.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B, 60,* 23–56.

Potthoff, R. F., Woodbury, M. A., & Manton, K. G. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association, 87,* 383–396.

Rao, J. N. K., & Thomas, D. R. (1989). Chi-square test for contingency tables. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 89–114). New York: Wiley.

Raudenbush, S., Bryk, A., & Congdon, R. T. (2002). *HLM 5.05*. Chicago: Scientific Software International.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association,* 308–313.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2,* 110–114.

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J.Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 59–87). New York: Wiley.

Stapleton, L. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling, 9,* 475–502.

Yuan, K., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30,* 167–202.

APPENDIX A
M*PLUS*

The M*plus* input file is as follows.

```
DATA: FILE = data\_file\_list.dat;
TYPE = MONTECARLO;
VARIABLE: NAMES ARE y1-y6 w;
MISSING = all(999);
WEIGHT = w;
ANALYSIS: estimator = mlr; type = missing h1;
MODEL:
[y1-y6@0];
y1-y6*1 (1);
i s | y1@0 y2@1 y3@2 y4@3 y5@4 y6@5;
i with s*0.3; i*1 s*0.2; [i*0.5 s*0.1];
```

The file data_file_list.dat is simply a list of the names of the data files.

```
Mplus\_1.dat
Mplus\_2.dat
…
Mplus\_500.dat
```


APPENDIX B
HLM

The first HLM input file is hlm.rsp.

```
Y
n
1
2
(A5, F11.6, F4.1)
hlml1.dat
1
(A5, F11.6)
hlml2.dat
Y
X
W
n
n
y
n
W
hlm.ssm
y
```

The second HLM input file is cmd.hlm.

```
NUMIT:10000
STOPVAL:0.0000010000
NONLIN:n
LEVEL1:Y = INTRCPT1 + X + RANDOM
LEVEL2:INTRCPT1 = INTRCPT2 + RANDOM/
LEVEL2:X = INTRCPT2 + RANDOM/
RESFIL:N
HETEROL1VAR:n
ACCEL:5
LVR:N
LEV1OLS:0
MLF:y
HYPOTH:n
FIXTAU:3
CONSTRAIN:N
OUTPUT:hlm.out
TITLE:NO TITLE
```

The two files are run within DOS with the command lines

```
hlm2s -r hlm.rsp
hlm2s hlm.ssm cmd.hlm
```

which produce an intermediate file hlm.ssm containing the sufficient statistics. In addition to that, this procedure is repeated within the DOS environment so that multiple datasets are analyzed.

APPENDIX C
MLwiN

As a first step, a large worksheet is allocated and all datasets are entered into it. The second step is to visually construct the model. The third step is to run the following macro file that essentially manages the multiple datasets, runs the estimation, and saves the results from each. The beginning of the macro file is as follows.

```
batc 1
MAXI 1000
calc c3 = c101
calc c5 = c102
weig
start
calc c1101 = c96
calc c1102 = c97
calc c1103 = c98
calc c1104 = c99
…
```