Finite Population Mixed Model Methods: New Developments

2. SPECIFIC AIMS

This research develops inferential methods using nonparametric finite population mixed models built directly from random variables used to formalize sampling schemes (Stanek and Singer 2004, Stanek, Singer et al. 2004). In this context, these random variables are embedded in a modeling framework similar to that commonly used in model-based approaches (Verbeke and Molenberghs 2000, McCulloch and Searle 2001, Raudenbush and Bryk 2002, Jiang 2007). The proposed research builds upon ongoing work by our research team in several areas yielding more accurate estimators/predictors with clear interpretation that impact applications in a broad range of health studies. In each area, the research proceeds by

- **Identifying a problem** and proposed solutions from the literature; conceptualizing the problem in a physical finite population setting; formally defining the population parameters, the sampling scheme, and measurement error random variables; and specifying their first and second central moments;
- **Developing estimators/predictors** for target parameters/target random variables;
- **Evaluating the performance relative to competitors** of the new estimators/predictors assuming known variance components;
- **Evaluating the performance of empirical estimators** obtained by substituting variance components by estimated ones via simulation; and
- **Developing software** (when none exists) to make implementation of the results practical in analysis.

As part of this process, our research team writes manuscripts for the peer reviewed literature presenting the new methods, and applications in practical contexts. The proposed research will extend current work through these stages of development in the following areas:

- **Accounting for endogenous and exogenous measurement error** including the development of expanded finite population models and predictors for such settings,
- **Development of guidance for practitioners** to identify settings where increased (or decreased) accuracy will result from accounting for auxiliary variables and confounding,
- **Extension of models and predictors** to settings where auxiliary variables include measurement error
- **Development of applications of auxiliary variables** suitable for use in predicting domain means and while accounting for other auxiliary variables as in synthetic estimation
- **Incorporation of cluster-level covariables** to develop predictors for targets corresponding to realized adjusted cluster means, measures of association, or regression coefficients
- **Extension to longitudinal study settings**, including growth curve problems
- **Extension to broader classes of experimental designs** including group randomized trials
- **Investigation of applications with missing data**
- **Extension to non-linear models**, including the development of applications of two-stage cluster sampling settings for prediction of realized cluster means when response is categorical

Based on our experience (see preliminary studies/progress), this process will result in new insights that will impact statistical practice, lead to clearer interpretation of statistical results, and foster new methods of nonparametric analysis applicable to a wide range of problems in public health and health sciences.

3. BACKGROUND AND SIGNIFICANCE

3.1 Significance

This research will impact inference for finite population studies where mixed models are usually applied, and will extend applications of such models to settings where their use is not typical. The impact can be traced to conceptual differences between the finite population mixed model, and alternative models, as discussed in the Preliminary Studies/Progress Section 4. Under finite population mixed model methods, unlike usual mixed models, only unobserved units, or units measured with error are predicted. In contrast to superpopulation models (Cassel, Särndal et al. 1977, Bolfarine and Zacks 1992), random variables in the finite population
mixed model are directly related to finite population parameter definitions and the corresponding distributions stem directly from the sampling scheme. Because the methods do not require parametric assumptions, they can be applied in practical settings where the distribution of subject characteristics is not known or specified. Also, unlike superpopulation model methods, the finite population mixed model methods can distinguish response error from nested unit effects (Stanek and Singer 2004). Advantages include the ability to clearly define target quantities (i.e., parameters/random variables) of interest directly relevant to the population being studied, the minimal need for assumptions, and the specification of a systematic procedure to derive estimators and predictors that typically outperform their competitors. As a consequence, we have improved accuracy of estimators/predictors, and clear interpretation of results.

This research will significantly extend our ability to define models specific to different physical settings, and to develop inference based on optimal methods that account for these settings. For example, in a cluster randomized trial designed to alter exercise and nutrition of employees at worksites, we will develop analysis methods that account explicitly for differences in worksite size, and can adjust treatment comparisons for known age, gender, and work history variables that are available on all employees based on clustered samples of workers. The analysis methods will formally incorporate important sources of uncertainty including subject specific endogenous measurement error (e.g., daily exercise variability), as well as exogenous measurement error (e.g., interviewer error), while separating and identifying variability due to different regression coefficients over time, and accounting for missing data, either intentional or unintentional.

The importance of this development is the enhanced ability it will convey to researchers attempting to understand complex real problems, identify the relative importance of uncertainty due to different sources, and present accurate results based on available data. A significant impact on practice is the clarity provided in predicting realized random effects, whether the effects correspond to treatment group comparisons, or subject specific changes in response. The significance of the approach is its comprehensive inclusion of purposeful study design decisions (sampling and experimental) as well as an analytic framework that seamlessly connects estimation and prediction to physically defined parameters. This research will extend the general methods in many directions, each of which expand our understanding of the relative contributions of uncertainty to inference, and develop software to implement the methods. The development is applicable in very large population settings where the notion of a finite population is valuable for problem definition and model development. Finally, this research will significantly improve understanding of competing statistical methods and frameworks since it directly links identifiable units (such as clusters, treatment groups, and subjects) to inference based on a linear model framework in a modern mixed model context.

3.2 Background

The background for this research was developed as part of an NIH-funded study entitled "Improving Analysis Methods for Cluster Randomized Prevention Trials" from 9/97-7/02 (NIH/NHLBI R01-HD36848, Stanek, PI) and a subsequent study entitled "Mixed Models for Finite Populations" from 4/1/05-3/30/08 (NIH/NHLBI R01-HL071828). The proposed research will include three investigators (Drs. Stanek, Singer, Li) from these previous studies, and take advantage of additional established collaborations.

This research will develop and extend finite population mixed model methods. In order to appreciate the importance of this research, it is helpful to consider the strengths and limitations of alternative model-based and design based approaches for inference. Under a model-based approach, flexible data analysis strategies for a variety of problems can rely on parametric linear models (Searle, Casella et al. 1992), generalized linear models (McCulloch and Searle 2001, Demidenko 2004), and superpopulation models (Valliant, Dorfman et al. 2000). In these settings, the observed data are usually assumed to represent a simple random sample, but the actual sampling scheme is rarely directly used to develop inference. The model-based approach carries some disadvantages, including the need for assumptions that can not be confirmed, the disregard for potentially useful known information on the population structure and how the data are obtained, the disconnection between the parameters in the stochastic model and the parameters in the population, and the separation of the statistical inference from the physical problem. Alternatively, design-based approaches connect the target parameters to the physical population, and can provide a robust basis for inference due to the need for minimal assumptions (Cochran 1977, Hedayat and Sinha 1991, Mukhopadhyay 2001). However, inferential methods using a design-based approach have focused on descriptive statistics such as the mean or total, and have not
provided the flexibility or appeal of the competing model-based approaches. Although intimately connected to the foundation for classical experimental design (Kempthorne 1955, Kruskal 1991, Hinkelmann and Kempthorne 1994), the actual sampling schemes have not been used as an inferential basis without additional model assumptions in studies involving regression and longitudinal settings. Combined model-assisted or calibration approaches (Särndal, Swenson et al., 1992, Brewer 2002) mix the advantages and disadvantages of each approach while failing to provide a firm framework for more complex problems.

The proposed research will significantly extend finite population mixed model methods (Stanek and Singer 2004, Stanek, Singer et al. 2004, Lencina, Singer et al., 2005, Stanek and Singer 2008, Li, Stanek et al. 2008, San Martino, Singer et al. 2008), taking advantage of the specificity of the design-based framework while retaining the flexibility of the model-based approaches. For background, we review the basic framework that enables sampling random variables to be formally integrated into linear models that is key to this research, and summarize the prediction theory that underlies the methodological work.

3.21 Using Sampling to Develop Inference: Basic Relationship Between Traditional Design Based Approach and the Finite Population Mixed Model Framework

The approach for representing selection of a simple random sample without replacement (srs) from a finite population dates back to Neyman (1934) in which the response for the \( i \)th sample subject is represented by

\[
Y_i = \sum_{s=1}^{N} U_{is} y_s ,
\]

with \( U_{is} \) denoting an indicator random variable that takes on a value of one if subject \( s \) is the \( i \)th selected subject and is zero otherwise, and \( y_s \) represents the non-stochastic response of subject \( s \) in a population of labeled subjects, \( s = 1, \ldots, N \). The random variable \( Y_i \) is indexed by the position in the sample, i.e. \( i = 1, \ldots, n \), and not by the subject’s label, \( s = 1, \ldots, N \). Once a subject is selected, the identity of the subject is realized. Explicitly, letting \( u_{is} \) represent the realized value of \( U_{is} \), the realized response for a subject can be expressed as

\[
\mu_s + \sum_{i=1}^{N} u_{is} \beta_s ,
\]

where \( \mu_s = \frac{1}{N} \sum_{i=1}^{N} y_s \) and \( \beta_s = y_s - \mu_s \). This suggests a direct connection between the representation of a randomly selected subject in the survey sampling literature, and the terminology used to describe realized random effects. In this context, we can directly represent response for the \( i \)th selected subject as

\[
Y_i = \mu_s + b_i \quad (1)
\]

where \( b_i = \sum_{s=1}^{N} u_{is} \beta_s \) is a random effect commonly used in mixed models, with its realized value given by \( \sum_{s=1}^{N} u_{is} \beta_s \). The term represented by \( b_i \) is random due to the selection of a subject, and is equal to the difference between the selected subject’s response and the mean response. The indicator random variables underlying the model have a long history both in sampling and in the experimental design literature (Hansen, Hurwitz et al. 1953, Cochran 1977, Kempthorne 1952, Kirk 1995), but their explicit connection to subjects in the population is unusual, largely due to difference between the timing of their development and timing of developments that popularized mixed models and prediction of realized random effects (Wilk and Kempthorne 1955, Scheffe 1956a,b, Dempster, Laird et al. 1977, Harville 1978, Laird and Ware 1982, Robinson 1991)). Using this connection, the random variables that formalize the sampling scheme can be directly employed to make inference in nonparametric models.

First, we discuss how the basic sampling random variables have been used in the traditional finite population sampling methodology, and how, with slight changes, they can be used to mimic a more standard stochastic linear model, as developed by Stanek, Singer et al. (2004). We consider a simple setting where we represent sample responses by the first \( n(=2) \) values in a permutation of \( N(=3) \) subjects in a population. Permutations re-order subjects, so the position in a permutation matters. We can formally represent random variables corresponding to a permutation of response as a sum of more elementary random variables, namely
\[
Y = \left( \begin{array}{c}
Y_1 \\
Y_2 \\
Y_3 \\
\end{array} \right) = \left( \begin{array}{ccc}
Y_{11} & Y_{12} & Y_{13} \\
Y_{21} & Y_{22} & Y_{23} \\
Y_{31} & Y_{32} & Y_{33} \\
\end{array} \right) = Y, \quad (Y) \quad (2)
\]

where rows correspond to positions, columns correspond to subjects, \( Y_i \) represents an \( N \times 1 \) column vector of ones, and the sample is represented by \( Y_i = (Y_1 \ Y_2)' \).

Traditional survey sampling (Cochran 1977, Särndal, Swensson et al. 1992, Thompson 1997, Brewer 2002)) summarizes the basic random variables in a different manner. First, notice that the sum of each column in (2) is the response of the corresponding subject,

\[
\begin{bmatrix}
Y_{11} & Y_{12} & Y_{13} \\
Y_{21} & Y_{22} & Y_{23} \\
Y_{31} & Y_{32} & Y_{33} \\
\end{bmatrix}' = \left( \begin{array}{c}
Y_1 \\
Y_2 \\
Y_3 \\
\end{array} \right) \quad (y) \quad (3)
\]

while the sum of the indicator random variables,

\[
U = \begin{bmatrix}
U_{11} & U_{12} & U_{13} \\
U_{21} & U_{22} & U_{23} \\
U_{31} & U_{32} & U_{33} \\
\end{bmatrix} \quad (4)
\]

over the sample in each column

\[
(1_n U_{11} U_{12} U_{13} \\
1_n U_{21} U_{22} U_{23} \\
1_n U_{31} U_{32} U_{33}) = \left( \begin{array}{c}
\sum_{i=1}^{n} U_{1i} \\
\sum_{i=1}^{n} U_{2i} \\
\sum_{i=1}^{n} U_{3i} \\
\end{array} \right) \quad (5)
\]

results in different random variables, the expectation of which corresponds to the inclusion probability for each subject. Representing \( E\left( \sum_{i=1}^{n} U_{1i} \right) = \pi_s \) as the inclusion probability for subject \( s \), the traditional finite population sampling approach represents the population via the pairs of values, \( (y_s, \pi_s) \), and a sample as in Table 1.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Inclusion Probability</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>s = 1</td>
<td>( \pi_1 )</td>
<td>( y_1 )</td>
</tr>
<tr>
<td>s = 2</td>
<td>( \pi_2 )</td>
<td>( y_2 )</td>
</tr>
<tr>
<td>s = 3</td>
<td>( \pi_3 )</td>
<td>( y_3 )</td>
</tr>
</tbody>
</table>

Table 1. Representation of a Sample of \( n = 2 \) using Traditional Finite Population Methods for a Realized Sample

This representation of the problem characterizes the design-based approach widely used in survey sampling (Cochran 1977, Särndal, Swensson et al., 1992), but its translation into a more standard stochastic linear model has not been considered without additional assumptions. One strategy to do so is the superpopulation model approach (Bolfarine and Zacks 1992, Valliant, Dorfman et al., 2000), where for each subject in Table 1, we assume that the response is a realization of a random variable, \( Y_s \), and we specify the expected value and variance for the joint set of random variables \( Y_s \), \( s = 1,...,N \). The sample corresponds to a set of \( n \) random variables that will be realized. This same framework is used to express intentionally missing data by Rubin 1976. An advantage of the superpopulation approach is the ease with which general linear models can be specified. However, inference is based on the superpopulation model assumptions, not the sampling random variables. Estimators motivated by the superpopulation models have been adapted to the design-based approach resulting in a strategy referred to as a model-assisted approach (Särndal, Swenson et al., 1992). Properties of such estimators, as the generalized regression estimator (GREG), for example, are evaluated with respect to the design, using the model only for motivation. Other strategies have been proposed that calibrate auxiliary variables to known values (Deville and Särndal 1992, Brewer 1999). While these approaches incorporate sampling (at least partially), they do not provide a general framework for extension to problems where model-based approaches are used, and may depend on model assumptions. We propose a
robust approach that makes use of the information on the structure of the population under study as well as the sampling scheme, and does not require model assumptions.

3.22 Estimation and Prediction Using the Finite Population Mixed Model

We review the basic framework for estimation and prediction in the finite population mixed model using the simplest setting to provide a context for developments and applications (Stanek, Singer et al. 2004). Our goal is to estimate finite population quantities (means, totals, measures of association, regression coefficients, population quantities standardized to national or state populations) or predict a linear combination of random variables (realized cluster mean, domain mean, realized treatment difference, realized subject mean) based on probability sample data obtained via simple random samples, stratified samples or two-stage cluster samples with (or without) error (endogenous or exogenous).

Several aspects of the formal statistical approach are appealing. First, we require the target parameter (or target random variable) to be clearly defined, and incorporate the definition explicitly in the model, facilitating interpretation of results. Sampling is represented explicitly via indicator variables. Second, we use a matrix representation of the model (Harville 1997) that parallels the standard approaches for linear mixed model, and facilitates the direct comparison. Third, we use a standard best linear unbiased prediction (BLUP) approach to develop optimal results (Royall 1968, 1970, 1973, 1986).

We illustrate these steps in a simple setting where the target parameter is the population mean, and we estimate it based on srs. We represent a model for a random permutation of the population similar to (2) by

\[ Y = U Y, \]  

and a simple random sample (without replacement) of \( Y \) by \( Y_i \), which, without loss of generality we take as the first \( n \) subjects in \( Y \). The sample is obtained by partitioning

\[ U = \begin{pmatrix} U_1' & U_2' \end{pmatrix} \]  

(7)

(where \( U_1 \) is \( n \times N \) and \( U_2 \) is \( (N-n) \times N \)) to obtain \( Y = \begin{pmatrix} Y_1' & Y_2' \end{pmatrix} \). We refer to a subject in \( Y_i \) as a primary sampling unit (PSU), and represent the corresponding response by \( iY \), \( 1 \leq i \leq n \). Notice that \( i \) corresponds to the position of a random variable in the permutation, not the label of the subject for the PSU in the sample (which corresponds to the random variable given by \( \sum_{i=1}^{N} U_{is} \)). Using properties of the indicator random variables that represent a permutation, we obtain

\[ E_\zeta(U) = \frac{1}{N} I_N I_N' \]  

and \( \text{var}_\zeta(\text{vec}(U)) = P_N \otimes P_N \), \( P_N = P_{n,N} \) when \( n = N \) where \( P_{n,N} = I_n - \frac{1}{N} J_n \), \( I_N \) corresponds to an \( N \times N \) identity matrix, \( J_N = I_N I_N' \), \( \otimes \) is the Kronecker product that multiplies each element of the matrix on the left by the matrix on the right, \( \text{vec} \) is the column expansion operator (Harville 1997), and \( \zeta \) denotes expectation with respect to the sampling distribution.

Using properties of \( U \), we express the finite population model for PSU \( i \) by (1), where

\[ E_\zeta(Y_i) = \mu_y, \]

\[ \text{var}_\zeta(Y_i) = \frac{N-1}{N} \sigma_y^2, \]  

\[ \text{cov}_\zeta(Y_i, Y_j) = -\frac{1}{N} \sigma_y^2, \]  

and \( \sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu_y)^2 \). We summarize model (1) for the entire population by

\[ Y = X \mu_y + b, \]  

(8)

where \( X = I_N \), \( b = U \beta = (b_1, b_2, \ldots, b_N)' \), \( \beta = (\beta_1, \beta_2, \ldots, \beta_N)' \), and \( \text{var}_\zeta(Y) = \sigma_y^2 P_N \).

Notice that (8) has a standard linear model form, but includes the entire set of \( N \) random variables, instead of solely the \( n \) sample random variables, with each random variable based on the finite population sampling. Unlike the representation of random variables in model-based approaches and in the missing data literature, the subscript \( i \) in \( Y_i \) corresponds to the position of the variable in a permutation, not the realized subject. Thus, \( Y_i \) is random because we do not know which subject will be realized in position \( i \) in a permutation.
Our research focuses on developing the predictors of population quantities defined as linear combinations of the random variables, namely $T = g'Y$ (for fixed known $g$). We require the predictor to be a linear function of the sample random variables, to be unbiased, and to have minimum expected mean squared error (MSE). To obtain the predictor, we let

$$
\begin{bmatrix}
Y_i \\
Y_{il}
\end{bmatrix} =
\begin{bmatrix}
Y_{i} \\
X_{il}
\end{bmatrix} \mu_i +
\begin{bmatrix}
b_i \\
b_{li}
\end{bmatrix}
$$

where $X_i = 1_n$, $X_{il} = 1_{n-a}$, so that $T = g_i'Y_i + g_{il}'Y_{il}$, where $g' = (g_i, g_{il})'$. After sampling, $g_i'Y_i$ is observed so that estimating $T$ requires only the prediction of $g_{il}'Y_{il}$. Using the prediction theory of Royall (1973), the best linear unbiased predictor (BLUP) of $g_{il}'Y_{il}$ is given by

$$
\hat{T} = g_i'Y_i + g_{il}' \left[ X_{il} \hat{\alpha} + V_{il} V_{il}^{-1} (Y_i - X_i \hat{\alpha}) \right],
$$

where $\hat{\alpha} = (X_i'V_{ii}^{-1}X_i)^{-1}X_i'V_{ii}^{-1}Y_i$, $\text{var} \left( \begin{bmatrix} Y_i \\ Y_{il} \end{bmatrix} \right) = \begin{bmatrix} V_i & V_{il} \\ V_{il} & V_{ll} \end{bmatrix}$ with $V_i = \sigma^2_i P_{iN}$, $V_{il} = \sigma^2_i P_{i(N-a)}$, and

$$
V_{il} = V_{il} = -\sigma^2_i \frac{J}{N}.
$$

For example, to predict $T = \mu_j$, we define $g = \frac{1}{N}1_N$, $g_i = \frac{1}{N}1_n$, and $g_{il} = \frac{1}{N}1_{n-a}$, resulting in

$$
\hat{T} = \frac{n}{N} \bar{Y} + \left( \frac{N-n}{N} \right) \bar{Y} = \bar{Y}
$$

which simplifies to the sample mean, $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

The fact that the best linear unbiased estimator under srs is the sample mean is not a new result. However, the strategy for directly incorporating sampling random variables to arrive at a linear model framework is new. Novel aspects are:

- while making use of the same underlying sampling indicator random variables as in the design-based framework, a standard linear model is defined, connecting sampling to contemporary modeling approaches.
- unlike the super-population model approach (that does make use of a standard linear model), the random variables are directly connected to the study population through the sampling.
- the representation of an elementary random variable is for a PSU in a permutation (corresponding to a position), not for a labeled subject. Because the subject for a PSU is random, the model is a simple mixed model.
- estimators of the population parameters are the sum of two terms, one corresponding to the contribution from the sample, and the other corresponding to a predictor of the contribution due to the remaining unobserved random variables. This interpretation is common in model-based prediction approaches, but novel in the context of approaches that only use random variables arising from the sample design.

4. PRELIMINARY STUDIES/PROGRESS REPORT

Our team has extended the finite population mixed model approach in several directions leading to new developments that build a solid foundation for translating the theoretical advances into practical applications. The general framework is flexible, and has been used recently to estimate population parameters, predict realized random effects, adjust for auxiliary covariables, confounding, and stratification, predict domain means and estimate measures of association, predict realized treatment differences in factorial designs, distinguish and account for interviewer error and subject response error, and predict latent values in two-stage cluster models with response error. Applications of these results address practical problems in previous and ongoing NIH-funded clinical and translational research projects and publicly-available health survey data (such as Behavioral Risk Factors Surveillance System (BRFSS)).

We summarize the progress made by the research team in: applications to srs with measurement error; problems accounting for auxiliary variables and confounding; prediction of domain means; estimation of measures of association and regression coefficients; prediction of random effects in two-stage sampling; and prediction of treatment contrasts in simple factorial experimental designs.
Past work has fostered the career development of five doctoral students. Three of them (Drs. Li, Lencina and San Martino) have graduated and now hold faculty appointments in research universities; each of them has developed his/her own expertise surrounding the central theme of finite mixed models and continues to make unique and independent contributions to this research. Two doctoral students (Ms. González and Xu) will graduate in the summer of 2008, each having their own distinct research directions. Two doctoral students (Mr. Moreno and Zhang) are expected to join the project after completion of their course work in the fall of 2008.

Extensions of the Finite Population Mixed Model

4.1 Measurement Error and Sampling

Measurement error is commonly incorporated in standard linear models, but rarely included in design-based models apart from experimental design settings. Measurement errors (also called errors of measurement by Cochran 1977, errors of observation by Sukhatme, Sukhatme et al. 1984 or response errors by Thompson 1997, Särndal, Swensson et al. 1992) occur when the recorded value for a subject differs from the true (latent) value. This topic has a broad literature, as discussed by Groves (1991). Two sources of measurement error can be identified; the first is related to the inherent natural variability of the subject (Kempthorne 1952, Hinkelman and Kempthorne 1994, Buonaccorsi 2006); the second is associated with the conditions of the measurement, such as the instrument, interviewer, or other aspects of the physical measurement process (Koch 1973, Fuller 1988, 1995). To clearly differentiate between the two types of measurement errors, we will term the first endogenous measurement errors and the second, exogenous measurement errors.

Stanek, Lencina et al. (2008) show that differentiating between these two sources of error has practical implications for predicting a subject’s latent value. For example, consider a model for saturated fat intake for subjects in the Seasons Study (Merriam, Ockene et al. 1999, Ockene, Chiriboga et al. 2004). On each study subject, telephone interviews were conducted on three days in each season by trained nutritionist interviewers. Endogenous error corresponds to daily variation in saturated fat intake, while exogenous error corresponds to error associated with interviewers assigned to selected days for sample subjects. Defining \( y_s \) as the true latent saturated fat intake over a season for subject \( s \), we represent saturated fat intake on the \( r \)th randomly selected day by

\[
Y_{sr} = y_s + E_{sr} = \mu_s + \beta_s + E_{sr}
\]

(10)

where \( E_R (E_{sr}) = 0 \) and \( \text{var}_R (E_{sr}) = \sigma_{sr}^2 \), with \( R \) denoting expectation with respect to the endogenous measurement error distribution. The random variable representing a response for the \( i \)th sample subject includes sample indicator random variables, \( Y_{ir} = \sum_{s=1}^{N} U_{is} Y_{sr} \), leading to a mixed model that includes endogenous measurement error with random subject effects given by

\[
Y_{ir} = \mu_i + b_i + E_{ir}^*\]

(11)

where \( E_{ir}^* = \sum_{s=1}^{N} U_{is} E_{sr} \) represents the deviation in saturated fat intake for the \( r \)th randomly selected day, \( r = 1, \ldots, R \) for the \( i \)th selected subject. Subsequently, we assume there is a single measure per subject (\( R = 1 \)) and drop the subscript \( r \), defining \( E_i^* = E_{ir}^* \) and \( Y_i^* = Y_{ir} \). We represent the model simultaneously for a permutation of subjects by including the indicator variable matrix \( U \) in the model, i.e.,

\[
U(y+E) = Uy + UE
\]

\[
Y^* = Y + E^*
\]

(12)

where \( E = \begin{pmatrix} E_1 & E_2 & \cdots & E_N \end{pmatrix} \) and \( E^* = UE = \begin{pmatrix} E_1^* & E_2^* & \cdots & E_N^* \end{pmatrix} \). Stanek, Lencina et al. (2008) derive the BLUP of the latent saturated fat intake for a realized sample subject as \( \hat{T}_i = \bar{Y}^* + k (Y_i^* - \bar{Y}^*) \), where \( k = \sigma_{i}^2 / (\sigma_{i}^2 + \bar{\sigma}_e^2) \) is the reliability ratio (Fuller 1988) and \( \bar{Y}^* = \frac{1}{N} \sum_{i=1}^{N} Y_i^* \). Note that \( \bar{\sigma}_e^2 = \sum_{i=1}^{N} \sigma_{sr}^2 / N \) in the coefficient \( k \) since
subjects are random effects in (11), and hence the average (over the population) endogenous measurement error variance is expected for a sample subject. In contrast, inclusion of exogenous measurement error arising from interviewers assigned to sample subjects will result in a model given by

$$
\hat{Y}_i = \mu + b_i + \tilde{E}_i \tag{13}
$$

where $b_i$ is defined following (1), $E_i(\tilde{E}_i) = 0$, and $\text{var}(\tilde{E}_i) = \sigma_i^2$. Assuming there is a single measure per subject, we represent the model simultaneously for a permutation of subjects by

$$
U\gamma + \tilde{E} = Y + \tilde{E}
$$

where $\tilde{E} = (\tilde{E}_1, \tilde{E}_2, \ldots, \tilde{E}_N)'$, where $\tilde{E}_i = \tilde{E}_i$. Under this model, the BLUP of the latent saturated fat intake for a realized sample subject is

$$
\hat{\mu} = \hat{\mu} + \frac{\sigma_i^2}{\sigma_i^2 + \sigma_i^2} (\hat{Y}_i - \hat{\mu})
$$

where $\hat{\mu} = \sum_{i=1}^n (\tilde{w}_i \hat{Y}_i / \sum_{i=1}^n \tilde{w}_i)$ is a weighted sample mean, $\tilde{w}_i = 1 / (\sigma_i^2 + \sigma_i^2)$, and $\hat{Y}_i = \hat{Y}_i$. While this predictor (which is equal to a weighted least squares predictor) is appropriate for exogenous measurement error, its use is not justified for heterogeneous endogenous measurement error. If used in such a setting, the predictor is biased, as discussed by Stanek, Lencina et al., (2008).

4.2 Accounting for Auxiliary Variables: Confounding

Estimates of basic epidemiological parameters, such as smoking and obesity prevalence, or the association between body mass index (BMI) and leisure time physical activity (in metabolic equivalent or METS) are usually obtained from sample surveys such as the BRFSS (Li, Kelsey et al., 2008, Kruger, Yore et al. 2007, Blanck, Dietz et al. 2006). The accuracy of such estimates can be improved by adjustment for auxiliary variables, such as gender or age groupings, and essentially involves weighting stratum specific estimates by population stratum weights. When data are collected by srs or stratified srs, such adjustment procedures correspond to post-stratification. For example, to control for gender when estimating smoking prevalence, $\pi_y$, based on a simple random sample, the post-stratified estimator (i.e. direct standardization) is given by

$$
\hat{\pi}_y = \sum_{g=1}^2 w_g \hat{P}_y, \quad \text{where } w,g \text{ is the proportion of males } (g=1) \text{ or females } (g=2) \text{ in the population, and } \hat{P}_y \text{ is the gender-specific sample smoking prevalence estimate.}
$$

In many applications, the population weights, $w_g$, are known, and the post-stratified estimator is recommended because it is considered to be more accurate than the crude sample prevalence, $\hat{\pi}_y = \sum_{g=1}^2 \left( n_g \hat{P}_y / \sum_{g=1}^2 n_g \right)$ where $n_g$ is the number of fe/males in the sample. Even though weights are known, the post-stratified estimator may be less accurate, as illustrated by Li, Kelsey et al. (2008) using a general model that accounts for auxiliary variables.

A general strategy for accounting for auxiliary variables in a srs or stratified sampling setting has been developed by Li (2003) and Li, Stanek, et al. (2008b). Let $z_s = \left( y_s, \delta_{q,s} \right)'$ denote a vector representing response and a set of $q=1,\ldots,Q$ auxiliary variables, $\delta_{q} = \left( \delta_{q1}, \delta_{q2}, \ldots, \delta_{qy} \right)'$ (centered at their known population mean, i.e., $\delta_{qs} = x_{qs} - \mu_{qs}$) for subject $s$ in a finite population of $s=1,\ldots,N$ subjects, where response for the population is summarized in the matrix $z = (y \quad x)',$ where $x = (\delta_1, \delta_2, \ldots, \delta_Q)'$ and $\delta_q = (\delta_{q1}, \delta_{q2}, \ldots, \delta_{qy})'$. The population mean and variance of $z$ are given by $\mu_z = (\mu_y, 0_y)'$ and $\frac{N-1}{N} \Sigma$, where,

$$
\Sigma = \frac{1}{N-1} \sum_{s=1}^N \left( z_s - \mu_z \right) \left( z_s - \mu_z \right)' = \begin{pmatrix} \sigma_y^2 & \sigma_{yv} \\ \sigma_{vy} & \sigma_{xx} \end{pmatrix}, \tag{14}
$$
\[ \sigma_{yx} \left( \begin{array}{cccc} \sigma_{y_1} & \ldots & \sigma_{y_q} \end{array} \right), \quad \sigma_{xx} = \left( \begin{array}{c} \sigma_{x_1x_1} \end{array} \right) \), and \[ \sigma_{x_q} = \sigma_{x_1y_q} \]. To introduce sampling, we permute the rows of \( z \) by pre-multiplying by (7), i.e., \( Z = Uz \), which we can represent by the multivariate model,

\[ Z = 1_n \mu' + (b \ d) \]

where \( b \) is defined following (8) and \( d = UX_\delta \). Using (7), we partition \( Z \) into vectors representing the sample and remaining random variables resulting in

\[ \begin{bmatrix} Z_t \\ Z_{II} \end{bmatrix} = \begin{bmatrix} \text{vec} \left[ U_I \left( y \ x_\delta \right) \right] \\ \text{vec} \left[ U_{II} \left( y \ x_\delta \right) \right] \end{bmatrix} \]

where

\[ V = \text{var}_{x_I} \left( \begin{bmatrix} Z_t \\ Z_{II} \end{bmatrix} = \begin{bmatrix} V_I \\ V_{II,I} \\ V_{II} \end{bmatrix} \right), \]

\[ V_I = \Sigma \otimes P_{n,N} \quad V_{II} = \Sigma \otimes P_{N-n,N} \quad \text{and} \quad V_{II,I} = -\Sigma \otimes \frac{1}{N-n(N-n)} J \quad \text{and} \quad \Sigma \text{ is given by (14)}. \]

A target parameter can be expressed as the linear combination

\[ T = c'Zq = \mathbf{g}'_{II}Z_{II} \quad \text{where} \quad \mathbf{g}_I = \mathbf{q} \otimes \mathbf{c}_I \quad \mathbf{g}_{II} = \mathbf{q} \otimes \mathbf{c}_{II} \quad \mathbf{c}_I = \left( c_1 \ c_2 \ \ldots \ c_n \right)', \quad \mathbf{c}_{II} = \left( c_{n+1} \ c_{n+2} \ \ldots \ c_N \right)' \quad \text{and} \quad \mathbf{c}_j = \left( c_1 \ c_2 \ \ldots \ c_n \right)' \].

When \( T = \mu_I \) (corresponding to smoking prevalence), \( c = \frac{1}{N}1_N \) and \( q = \left( 1 \ 0 \_Q \right)' \). Because \( Z_I \) will be realized after selecting the sample, estimating \( T = \mu_I \) simplifies to predicting the remaining linear combination of the unobserved random variables in \( Z_{II} \), namely \( \mathbf{g}_I^T Z_{II} \). Using the prediction theory of Section 3.2, the BLUP of \( \mathbf{g}_I^T Z_{II} \) corresponds to \( \left( N - n \right) \left( \bar{Y} - \frac{N}{N-n} \sum_{q=1}^Q \gamma_q \bar{d}_{qI} \right) \) where \( \gamma_q = \sigma_{y_q} / \sigma_{x_q}^2 \) and \( \bar{d}_{qI} = \frac{1}{n} \sum_{i=1}^n d_{qi} \). When added to the realized sample portion of \( \mu_I \), the BLUP is given by

\[ \hat{T} = \frac{n}{N} \bar{Y} + \left( \frac{N-n}{N} \right) \left( \bar{Y} - \frac{N}{N-n} \sum_{q=1}^Q \gamma_q \bar{d}_{qI} \right) \].

Expressed in this form, the predictor of the unobserved (intentionally missing) response, \( \mathbf{g}_I^T Z_{II} \), are adjusted by the auxiliary variables, leading to an estimator with \( \text{var} \left( \hat{T} \right) = \left( 1 - \rho^2_{yx} \right) \left( \frac{N-n}{N} \right) \sigma^2_y \), where \( \rho^2_{yx} = \sigma^2_{yx} / \sigma^2_{x_q} \) is the squared multiple correlation coefficient. When values of \( \gamma_q \) are known, the MSE of (18) is smaller than the MSE of \( \bar{Y} \) by \( 1 - \rho^2_{yx} \), and is equal to that of the post-stratified estimator, \( \hat{T} = \hat{P}_{ps} \).

Accounting for auxiliary variables reduces the MSE and, when interpreted conditionally on the sample, a common practice in Epidemiology (Greenland and Robbins 1986, 1999), removes bias. These advantages come at the price of knowledge of \( \gamma_q \), \( q = 1, \ldots, Q \), a fact made clear in (18). Although in some settings, auxiliary information is known for the entire population so that \( \sigma_{yx} \) is known, \( \sigma_{yx} \) must be estimated from the sample or additional data. Li, Stanek et al., (2008b) study the properties of the empirical estimator formed by replacing \( \sigma_{yx} \) by a sample estimate, \( \hat{\sigma}_{yx} \) in (18) via simulation. For example, when estimating smoking prevalence adjusting for gender based on a sample of size 100, the results indicate that post-stratification will improve accuracy only when the relative smoking prevalence in males is more than 1.6 times the smoking prevalence in females, assuming equal proportions of males and females in the population. Because such large differences in stratum specific smoking prevalence rates are rare, Li, Stanek et al., (2008b) conclude that contrary to common practice, direct standardization often should not be done because the empirical estimators are less accurate than the simple unadjusted mean. Additional work is planned in this proposal to extend such
practical guidance to other settings, such as for continuous or multinomial variables, and auxiliary variables with measurement error (including mis-classification).

4.3 Domains

Estimates for domains (i.e., small areas, gender groups, age groups, or a nursing unit within a hospital) are frequently of interest in finite population settings (Jiang and Lahiri 2006a,b). Li (2008) and Li, Kelsey, et al. (2008b) have investigated smoking and obesity prevalence in communities in Massachusetts using the BRFSS to identify high-priority locales that can be targeted for intervention. A distinguishing feature of this problem is the inability to identify a sample unit’s domain membership until the units are realized in the sample. Domain membership, such as respondent’s town/city and 5-digit zip code, has been included in the Massachusetts BRFSS since 1999 along with other responses such as self-reported smoking status, body weight, height, and other demographics. Through aggregation of census tract or census block group-level data from the US Census 2000, or annual official lists of residents in municipalities (commonly called “Town List”) in the Commonwealth of Massachusetts, population data on demographic variables for small areas can be constructed. Such data yield aggregate known auxiliary information on individual domains, offering the potential for auxiliary adjustment when obtaining small area estimates, similar in spirit to the superpopulation model approach developed by Holt, Smith et al. (1979). Mixed models motivated through similar superpopulation models are commonly recommended (Rao 1999). Li, Land et al. (2008) and Li, Kelsey, et al. (2008b) have constructed small area estimation models for body mass index (BMI) (continuous or binary) using conventional mixed models, and efforts have been made to adapt finite population mixed models for this purpose.

Methods for domain estimation using finite population sampling models have been developed by Lencina, Singer et al. (2007) who conditioned on the realized domain membership and sample size to estimate domain characteristics in a stratified finite population. For example, assume that census data is used to obtain population age/gender totals in zip-code areas based on recent census data, and random digit dialing is used to interview a sample of adults in a given state. We label age/gender strata by \( h = 1, \ldots, H \), and label domains (corresponding to zip-code areas) by \( i = 1, \ldots, I \). Let us represent response for subject \( t = 1, \ldots, N_h \) in stratum \( h \) by \( y_{ht} \), defining \( \mathbf{y}_h \) as the \( N_h \times 1 \) potential response vector for stratum \( h \). Defining \( \mathbf{U}_h \) as an \( N_h \times N_h \) matrix of indicator random variables similar to (4), we represent a random permutation of subjects in stratum \( h \) by \( \mathbf{Y}_h = \mathbf{U}_h \mathbf{y}_h \), concatenating strata into a population vector given by \( \mathbf{Y} = \left(\mathbf{Y}_1^\prime \quad \mathbf{Y}_2^\prime \quad \cdots \quad \mathbf{Y}_H^\prime\right)^\prime \). This formally incorporates sampling into a stratified finite population setting. Defining \( \mu_h = \mathbf{1}_{N_h}^\prime \mathbf{y}_h / N_h \) and \( \sigma_h^2 = \mathbf{y}_h^\prime \mathbf{P}_{N_h} \mathbf{y}_h / (N_h - 1) \) as stratum parameters, \( E(\mathbf{Y}) = \left(\bigoplus_{h=1}^H \mathbf{1}_{N_h} \right) \mu \) while \( \text{var}(\mathbf{Y}) = \bigoplus_{h=1}^H \sigma_h^2 \mathbf{P}_{N_h} \), where \( \bigoplus \) represents an operator that places the arguments on the diagonal to form a block diagonal matrix. Conditionally on the number of sample subjects in a domain in each stratum, Lencina, Singer et al. (2007) derived estimators of domain means based on the conditional finite population mixed model. The estimator simplifies to a synthetic estimator (Marker 1999) when no sample elements are selected in a domain.

4.4 Association

In many settings, a measure of association, \( \alpha = \sum_{j=1}^N (y_j - \mu_j) (x_{qj} - \mu_{qj}) / \left(\sum_{j=1}^N (x_{qj} - \mu_{qj})^2\right) \) is of principal interest, as for example, the association between daily caloric intake (kcal/kg), \( y_j \) in kcal/kg, and body mass index (BMI), \( x_{qj} \) expressed as weight/height\(^2\). González, Singer et al. (2008) have developed nonparametric estimators of a measure of association based on sampling indicator random variables by introducing weights in response, where \( y_j = w_j y_j^\prime \) (with the superscript “dot” denoting un-weighted values). When \( w_j = \left(\sum_{j=1}^N (x_{qj} - \mu_{qj})^2\right)^{-1} \left[\sum_{j=1}^N (x_{qj} - \mu_{qj})^2\right] \), the weighted total, \( T = \sum_{j=1}^N y_j = \alpha \) is a measure of association of response with
Using the weighted response, the association is a linear combination of $Z$ given by (17) where $c = 1_N$, and $q = (1 \ 0_Q)'$, and $Q$ is the number of auxiliary variables. Since response is observed only for sample subjects, an estimate of $T$ is developed by predicting the response sum for subjects not in the sample, i.e. $g_l'Z_H$. The values of the auxiliary variable can reduce the MSE of the predictor when \[ \rho^2_{xy} = \frac{\alpha' \sigma_{xy} \alpha / \sigma^2_y}{\sigma^2_{xy}}, \] the squared multiple correlation is not zero.

In practice, when there is endogenous measurement error in response as in (10), even the latent response for $y$ is observed only for sample subjects, so that $\mathcal{Y} = \{y_{sk} : s = 1, \ldots, N_k, k = 1, \ldots, Q\}$ will be observed, not $y$. González, Singer et al. (2008) extends the predictor to this setting assuming that all endogenous measurement errors are independent, $E_R(E_{sk}) = 0$, $E_{sk} = w_{sk}E_{sk}$, $\var{R}(E_{sk}) = w_{sk}^2\sigma_{se}^2 = \sigma_{se}^2$ and $\var{R}(E_{sk}) = \sigma_{se}^2$. When there is one response per subject, the development extends $Z$ to $Z^* = Z + U(E_{0_{N-Q}})$ which we express as the multivariate mixed model

\[
Z^* = X(\mu) + I_{N} (b) + \left( E^* 0_{N=Q} \right),
\]

where $X = 1_N$, $b$ is given in (8) and $d_x$ is given in Section 4.2. Partitioning $Z^*$ into the sample and remainder,

\[
\begin{pmatrix} Z^*_I \\ Z^*_H \end{pmatrix} = \begin{pmatrix} \text{vec}[U_I(E 0_{N=Q})] \\ \text{vec}[U_H(E 0_{N=Q})] \end{pmatrix}
\]

where $V = \text{var}_R\begin{pmatrix} Z^*_I \\ Z^*_H \end{pmatrix} = \begin{pmatrix} V^*_I \\ V^*_II, V^*_H \end{pmatrix}$ which we represent as $V^* = V + V_R$ where $V$ is given by (16) and $V_R = \begin{pmatrix} V_{RI} \\ 0_{N-Q} \end{pmatrix}$ where $V_{RI} = \Sigma_R \otimes I_n$, $V_RR = \Sigma_R \otimes I_{N-n}$, and $\Sigma_R = \begin{pmatrix} \sigma^2_{e} & 0_Q \\ 0_Q & 0_{Q-Q} \end{pmatrix}$. Notice that sampling will result in observing the values of $Z^*_I$, while the measure of association is defined by (17) as the linear combination

$T = g_l'Z_I + g_l'H_Z_k$ with $c = 1_N$ and $q = (1 \ 0_Q)'$. In order to develop an unbiased predictor $\hat{T}$ that is a linear function of $Z^*$ and has minimum MSE, we use the joint distribution of $\begin{pmatrix} Z^*_I \\ Z^*_H \end{pmatrix}$.

\[
\hat{T} = n_{Q} \sum_{i} \left[ X_i \hat{\alpha} + V_{i}^{-1} \left( Z_i - X_i \hat{\alpha} \right) \right] + g_l' \left[ X \hat{\alpha} + V_{i}^{-1} V_{i}^{-1} \left( Z_i - X_i \hat{\alpha} \right) \right]
\]

where $\hat{\alpha} = \left( X_i V_{i}^{-1} X_i \right)^{-1} X_i V_{i}^{-1} Z_i$, $X_i = 1_n$, and $X_H = 1_{N-n}$ which simplifies to (18) with the sample average weighted response $\bar{Y}$ replacing $Y$. When $Q > 1$, it is possible to develop similar measures of association that are adjusted for other auxiliary variables, similar to adjusting via post-stratification for confounding of the other variables. On-going research is underway to extend results to include endogenous measurement error in the auxiliary variables and integrate such results with other modeling approaches (Chen, Hong et al., 2005).

### 4.5 Two-Stage Cluster Sampling

In many settings, data arise from a finite population but information on the finite structure is not incorporated in the analysis. Our research demonstrates that if such information is not used in the analysis (such as in a typical mixed-model application), there is loss in accuracy (Stanek and Singer 2004, 2008, San Martino, Singer et al. 2008). We find a large loss when cluster sizes are small and measures are made on an appreciable fraction of the cluster members, such as with clusters of children in families, physician practices in hospitals, classrooms in schools, and carpal tunnels in one or two hands of subjects. Cluster size also can have an impact in clustered study designs such as the Watch Study (NIH/NHLBI 2R18 HL4492. PI: Ockene
“Systems to Enhance Provider Counseling in Hyperlipidemia” (Ockene, Hebert, et al., 1996) to evaluate a nutritional and exercise intervention in Worcester, MA, and studies in the Worksite Intervention Program, to evaluate the impact of a worksite intervention on weight gain (NIH/NHLBI 5R01HL079483 PI: Lemon, “Active Living and Healthy Diet at the Workplace”).

Stanek and Singer (2004) have developed nonparametric mixed models for two-stage cluster sampling with endogenous measurement error in populations where clusters are of equal size and equal size samples are selected per cluster. They used the models to predict random effects corresponding to cluster latent values. For example, defining a physician’s practice as a cluster of patients, and a single measure of serum cholesterol on a patient as response, predictors were obtained for the latent serum cholesterol value for a practice. Such predictors are important in developing effective practice-based dietary and physical activity interventions to alter patient’s cholesterol, as in the Watch II study (Ockene, Chiriboga et al., 2008). We outline their results beginning with the endogenous measurement error model similar to (10), and representing the \( r^{th} \) response for subject \( i \) in cluster \( s \) by

\[
Y_{is} = y_{is} + W_{is} = \mu + \beta_i + \epsilon_{is} + W_{is}
\]

(21)

where \( \mu = \mu_j \), \( y_s = \frac{1}{M} \sum_{i=1}^{M} y_{is} \), \( \epsilon_{is} = y_{is} - y_s \), and \( W_{is} \) represents endogenous measurement error. In (21), \( t = 1, ..., M \) labels subjects in cluster \( s \), \( s = 1, ..., N \), and we assume for \( r = 1, ..., R \) that \( E_R(W_{is}) = 0 \) and

\[
\text{cov}_R(W_{is}, W_{js}) = \sigma_{is}^2 \quad \text{if } str = s^r \quad \text{or zero otherwise.}
\]

We define the variance of subject latent values in cluster \( s \) by \( \frac{1}{M} \sum_{i=1}^{M} (y_{is} - y_s)^2 \), and the variance of cluster latent values by \( \frac{N-1}{N} \sum (y_s - \mu)^2 \).

We represent the \( NM \) values in the population by \( y = (y_1', y_2', \ldots, y_N') \) where \( y_s = (y_{s1}, y_{s2}, \ldots, y_{sM})' \).

We define random variables that represent a two-stage random permutation of the population, and use them to develop a predictor based on a two-stage cluster sample of \( m \) sample subjects in each of \( n \) sample clusters. To do so, we use the indicator random variables that permute clusters given by (4) and define a similar \( M \times M \) matrix of indicator random variables that permute subjects in a cluster by \( U^{(i)} \) for \( s = 1, ..., N \), where

\[
U^{(i)} = (U_1^{(i)}, U_2^{(i)}, \ldots, U_M^{(i)}) \quad \text{and} \quad U^{(i)}_s = (U_1^{(i)} s, U_2^{(i)} s, \ldots, U_M^{(i)} s)'.
\]

The entire permuted population is given by

\[
Y = (U \otimes I_M)(\bigoplus_{s=1}^{N} U^{(s)}) y,
\]

where the random variable representing response for the subject in position \( j \) (which we refer to as the \( j^{th} \) secondary sampling unit (SSU \( j \))) in the cluster in position \( i \) (PSU \( i \)) is given by

\[
Y_{ij} = \mu_j + b_i + \epsilon_{ij}
\]

where \( Y_{ij} = \sum_{s=1}^{N} \sum_{i=1}^{M} U_{is} U_{ij}^{(i)} y_{st} \), \( b_i \) is defined following (1) and \( \epsilon_{ij} = \sum_{s=1}^{N} \sum_{i=1}^{M} U_{is} U_{ij}^{(i)} \epsilon_{is} \). Including

\[
W_{ij} = \sum_{s=1}^{N} \sum_{i=1}^{M} U_{is} U_{ij}^{(i)} W_{is},
\]

the finite population mixed model is given by

\[
Y_{ij} = \mu_j + b_i + \epsilon_{ij} + W_{ij}
\]

(22)

which we express simultaneously for all PSUs and SSUs in the population (when \( R = 1 \)) via the mixed model

\[
Y^* = X_{ij} \mu + Z_{ij} + e + W^*
\]

(23)

where \( X = I_{NM} \), \( Z = I_N \otimes I_M \), and where \( W^* \) and \( e \) are defined similarly to \( Y \), respectively. This can be expressed in terms of

\[
Y = X_{ij} \mu + Z_{ij} + e,
\]

(24)

where \( Y^* = Y + W^* \). The expected value of \( Y^* \) is given by \( E_{\xi_{ij}, \xi, \mu} (Y^*) = X_{ij} \mu \) where \( \xi_{ij} \) and \( \xi_j \) denote expectation with respect to the distribution of PSUs and SSUs, respectively, and the variance is

\[
\text{var}_{\xi_{ij}, \xi, \mu} (Y^*) = \langle \sigma_x^2 + \sigma_e^2 \rangle I_{NM} + \sigma_x^2 (I_N \otimes J_M) - \frac{\sigma_x^2}{N} J_{NM}
\]

where \( \sigma_x^2 = \sum_{s=1}^{N} \sum_{i=1}^{M} \sigma_{is}^2 J_{NM} \), \( \sigma_{is}^2 = \frac{1}{N} \sum_{s=1}^{N} \sigma_{is}^2 \), and \( \sigma^2 = \sigma^2 - \frac{\sigma_x^2}{M} \).
Parameters and random variables of interest are defined as linear combinations of the random variables of the form $Y = gY$ where $g_i = (g_{i1}, g_{i2}, \ldots, g_{in})$ with $g_i = (g_{im}, \ldots, g_{iM})'$, $i = 1, \ldots, N$, is a vector of known constants.

Linear combinations of latent values for PSUs may be defined by taking $g' = c' \otimes I_n / M$ where $c$ is a vector of known constants. Of principal interest is the linear combination of $Y$ that defines the latent value of PSU $i$, with $c = c_i$ denoting an $N \times 1$ vector with a value of one in position $i$, and zero elsewhere. Note that elements of $Y$ are observed as opposed to elements of $Y$.

The development of the BLUP of the latent value of a realized cluster in position $i \leq n$ is similar to that given in Section 4.4, with details presented by Stanek and Singer (2004). First, random variables in (23) and (24) are re-arranged into a sample and remainder to form $(Y_i^*, Y_j^*, Y_k^*)'$, where $\text{var}_{i \in \mathbb{R}} (Y_i^*) = Y_i + \sigma^2 Q_{nn}$, and $\text{var}_{i \in \mathbb{R}} (Y_j^*) = Y_j + \sigma^2 Q_{nn}$, and $\text{var}_{i \in \mathbb{R}} (Y_k^*) = Y_k + \sigma^2 Q_{nn}$.

The finite population models provide a seamless transition to problems where the cluster size (assumed to be very large) is not relevant.

<table>
<thead>
<tr>
<th>Table 2. Predictors of the Latent Value of PSU $i$ when $i \leq n$ in Two-stage Cluster Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
</tr>
<tr>
<td>Mixed Model</td>
</tr>
<tr>
<td>Scott&amp;Smith</td>
</tr>
<tr>
<td>Finite Pop. MM</td>
</tr>
<tr>
<td>Finite Pop MM</td>
</tr>
</tbody>
</table>

The results in Table 2 provide valuable insight for interpreting predictors of the latent value of a realized random effect. Each predictor is the weighted sum of a term predicting the latent values for the SSUs in the sample, and a term predicting the latent values for the remaining SSUs. The last three predictors, unlike the mixed model predictor, give weight to SSUs realized in the sample. The mixed model predictor can be seen to be the limit of Scott and Smith’s (1969) predictor as the size of the PSU increases. This finding implies that the finite population models provide a seamless transition to problems where the cluster size (assumed to be very large) is not relevant.

The predictor developed by Scott and Smith (1969) is nearly identical to the predictor based on the finite population mixed model, differing only in definitions of variance components and shrinkage constants. A key difference is the use of $\sigma^2_i$ instead of $\sigma^2$ in defining the shrinkage constants, $k_i$ and $k$. The finite population mixed model development reveals that the predictors from both Scott and Smith and the mixed model consider the $U_{is}$ in $T_i = \sum_{s=1}^{S} U_{is} \mu$, to be random when defining variances, but condition on these random variables when defining $k_i$. In this respect, the superpopulation model and mixed model theory is not consistent. The finite
population mixed model avoids this problem, and separates shrinkage due to endogenous measurement error, i.e., governed by $k^*$, from shrinkage due to random clusters and endogenous measurement error considered simultaneously, i.e., governed by $k$. The ability to separate these sources of variability is a direct result of the explicit inclusion of sampling random variables in the endogenous measurement error model.

The predictors defined in Table 2 have uniformly smaller MSE than the mixed model or Scott and Smith’s model predictors when variance components are known. Reduction in the MSE relative to the mixed model predictor are dramatic (over 50%) when $f \geq 0.9$ and the subject intra-class correlation, $\rho = \sigma^2_s / \left( \sigma^2_s + \sigma^2_e \right)$, is larger than 0.8. Similarly large reductions in the MSE occur relative to Scott and Smith’s predictor when $f \geq 0.9$ and the subject intra-class correlation is less than 0.4.

When variance components are not known, the properties of empirical predictors based on methods of moments estimates of variance components have been studied by San Martino, Singer et al. (2008) via an extensive simulation study involving 840 populations and 8904 settings. The results indicate that the MSE of the finite population mixed model predictors summarized in Table 2 is either the best or equivalent to the best predictor or equivalent to the best predictor (MSE within 15% of best) in 90-100% of the settings.

### 4.6 Two-Stage Cluster Sampling with Unequal Size Clusters

The finite population mixed model given by (24) was developed for equal size clusters. There are many settings where cluster sizes are different, as in hospital/organizational level assessment where there are different numbers of patients for a physician (cluster), and quality/performance measures are linked to reimbursement. An expanded finite population mixed model has recently been developed by Stanek and Singer (2008) for two-stage samples of unequal size clustered populations. Best linear unbiased predictors resulting from this model outperform (i.e., have smaller MSE) all the predictors in Table 2, even in equal cluster size settings. The new model represents response for different size clusters by an expanded set of random variables that retain the link between the latent values for clusters and PSUs. The need for the expansion can be best seen via an example. Suppose our interest is in the average cost of appendectomies (the latent value) for each of three hospitals in the past year (Table 3), and that such costs are known (without error) for some patients in two of the hospitals. When the data are obtained from a stratified srs of appendectomy patients, with hospitals as strata, the best linear unbiased estimate is the average cost for the available patients in each hospital (i.e., $2000 for Central, and $1800 for Mercy).

<table>
<thead>
<tr>
<th>Hospital</th>
<th>$M_i$</th>
<th>Mean</th>
<th>Variance</th>
<th>Patient* $(i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>s=1 (County)</td>
<td>2</td>
<td>$\mu_1$</td>
<td>$\sigma^2_1$</td>
<td>$y_{11}$</td>
</tr>
<tr>
<td>s=2 (Central)</td>
<td>4</td>
<td>$\mu_2$</td>
<td>$\sigma^2_2$</td>
<td>$y_{21}$</td>
</tr>
<tr>
<td>s=3 (Mercy)</td>
<td>2</td>
<td>$\mu_3$</td>
<td>$\sigma^2_3$</td>
<td>$y_{31}$</td>
</tr>
</tbody>
</table>

* Names are fictitious

Now assume that a srs of appendectomy patients is selected from each of a srs of hospitals (Table 4) according to a two-stage sampling scheme. We refer to a sample hospital as a primary sampling unit (PSU) to distinguish it from a specific hospital, and to a sample patient as a secondary sampling unit (SSU) to distinguish it from a specific patient. Under the usual mixed model (similar to (22) without endogenous measurement error), the BLUP of $T_i = \mu + b_i$ is given in Table 2, and does not use additional information, such as the number of hospitals in the population, or the number of appendectomy patients in each hospital, even though such additional information may be available (as illustrated by the representation of the remaining
random variables in Table 4). When clusters differ in size, use of such information has the potential to result in more accurate predictors.

A superpopulation model attempts to account for such additional information. The combined sample and remainder represent a superpopulation that is constructed by first (conceptually) selecting a finite population (presumably from some larger population in time or space), and then selecting a two-stage sample from it, leading to a predictor similar to $\hat{P}_i$ in Table 2 with $f_i$ (the sampling fraction for PSU $i$) replacing $f$. The superpopulation model predictor is developed assuming the random variables do not depend on the labeled clusters. This assumption is false, as can be seen by noting that hospital associated with the sample PSU $i = 1$ in Table 4 can not correspond to County Hospital in Table 3.

<table>
<thead>
<tr>
<th>Sample PSU (i)</th>
<th>Sample Size</th>
<th>Latent Value</th>
<th>Sample Patient</th>
<th>SSU (j)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i = 1$</td>
<td>3</td>
<td>$\mu + b_1$</td>
<td>$Y_{11}$</td>
<td>$Y_{13}$</td>
</tr>
<tr>
<td>$i = 2 = n$</td>
<td>2</td>
<td>$\mu + b_2$</td>
<td>$Y_{12}$</td>
<td>$Y_{13}$</td>
</tr>
<tr>
<td>$i = 1$ (Central)</td>
<td></td>
<td></td>
<td></td>
<td>$Y_{14}$</td>
</tr>
<tr>
<td>$i = 2$ (Mercy)</td>
<td></td>
<td></td>
<td></td>
<td>$Y_{23}$</td>
</tr>
<tr>
<td>$i = 3$ (County)</td>
<td></td>
<td></td>
<td>$Y_{31}$</td>
<td>$Y_{32}$</td>
</tr>
</tbody>
</table>

Stanek and Singer (2008) overcome the limitation of the mixed model and superpopulation model by defining an expanded set of random variables. Let the response for SSU $j$ in cluster $s$ be expressed as $\bar{Y}_s = \sum_{j=1}^{M_s} U_j^{(s)} Y_{j,s}$, and define the weighted response as $\bar{Y}_s = w_{s,j} \bar{Y}_s$ so that the sum, $\sum_{j=1}^{M_s} \bar{Y}_s$ will correspond to a cluster total when $w_{s,j} = 1$ for all $j = 1, ..., M_s$, or to a cluster mean when $w_{s,j} = 1 / M_s$ for all $j = 1, ..., M_s$. If all clusters were equal in size, we could represent a permutation of SSUs for PSU $i$ by $\sum_{s=1}^{N} U_i^{(s)}$, where the vector $\bar{Y}_s = \left( \bar{Y}_{s1}, \bar{Y}_{s2}, ..., \bar{Y}_{sM_s} \right)'$ represents a permutation of weighted responses for the SSUs in cluster $s$. When cluster sizes differ, this sum is not defined, since the dimensions of the vectors composing it cannot all be equal. We solve this problem by expanding the set of random variables similarly to (3) associated to PSU $i$ into the $N \times 1$ vector $\tilde{Y}_i = \left( \tilde{Y}_{i1}, \tilde{Y}_{i2}, ..., \tilde{Y}_{iN} \right)'$ so that a two-stage random permutation of the population is represented by the $N \times N$ vector, $\tilde{Y} = \left( \tilde{Y}_{11}, \tilde{Y}_{12}, ..., \tilde{Y}_{NN} \right)'$, where the $j^{th}$ element of $\tilde{Y}$ that corresponds to $U_{i,s}$ is $U_{i,s} \tilde{Y}_{i,j}$ and $N = \sum_{s=1}^{N} M_s$.

To obtain an optimal predictor of the latent value of PSU $i$, $T_i = \sum_{s=1}^{N} U_{i,s} \left( \sum_{j=1}^{M_s} \tilde{Y}_{i,j} \right)$, we adopt the basic strategy outlined in Section 3.22. While in theory, an optimal predictor can be obtained via this recipe, in practice, the high dimensionality of the expanded random vectors may result in singularities that lead to multiple solutions as discussed in Stanek, Singer et al. (2004). For this reason, Stanek and Singer (2008) identify projections of the expanded random variables onto a lower dimensional space that retain the necessary information for an optimal solution when $w_{i,j} = w_s$ for all $j = 1, ..., M_s$. The resulting BLUP for a linear combination of PSUs is given by
\[
\hat{T}_p = \sum_{i,j} c_i \left( \hat{Y}_{ij} \right) + \frac{N}{n} \left( \sum_{i,j} f_{ij} M_i \bar{Y}_{ij} \right) + \frac{N - n}{n} \left( \sum_{i,j} I_s(M_i f_{ij} \bar{Y}_{ij}) \right)
\]

where \( \hat{Y}_s = \frac{1}{n} \sum_{i,j} \hat{Y}_{ij} \), \( k_s = k \), \( d_s = \frac{1}{N} \sum_{i,j} \left( 1 - k_s \right) d_{ij} \), \( f_s = \frac{m_s}{M_s} \), \( d_s = M_s \mu_s \), \( k_s = \frac{f_s^2 d_s^2}{f_s^2 d_s^2 + (N - 1) v_{se}^2} \), \( k = \frac{1}{N} \sum_{i,j} c_i \), \( v_{se} = f_s (1 - f_s) \frac{M_s \sigma^2}{N} \), \( \bar{c}_s = \frac{1}{N} \sum_{i,j} c_i \), \( \bar{c} = \frac{1}{N} \sum_{i,j} c_i \) and \( I_s = \sum_{i,j} U_{ij} \) is an indicator ‘inclusion’ random variable for cluster \( s \) in the sample. Explicit expressions for the MSE of (25) are given by Stanek and Singer (2008). The predictor (25) outperform all predictors in Table 2, even when clusters are of equal size.

The expanded finite population mixed model uses a larger set of \( 2N^2 \) random variables than the \( 2N \) random variables typically used in superpopulation models or in the finite population mixed model of Stanek and Singer (2004). These random variables are fewer than the \( N^2 \) random variables resulting from an expansion that retains the identity of units and SSUs, and even fewer than the very general representation of the model used by Godambe (1955). The intermediate set of random variables allows a clear representation of a two-stage sample while accounting for details on different cluster and sample sizes, while other approaches do not appear to connect the potentially observable data to the random variables in the stochastic model.

Using a theorem proposed by Rao and Bellhouse (1978), Stanek and Singer (2008) illustrate how different finite population mixed models can be compared by considering them in a hierarchy and identifying whether the additional set of orthogonal random variables adds to the information about the target quantity. They show that further reductions in the number of random variables will lead to loss of information. The results also indicate that different target quantities may require different sets of random variables to be optimally predicted. For example, if there is interest in the relationship between two variables among units (in a cluster), the collapsed expanded set of \( 2N^2 \) random variables is not likely to be sufficient. Further work is planned to develop optimal predictors for such important problems.

4.7 Experimental Design

Randomized controlled trials (RCT) are often blinded as to the actual treatment administered in an effort to eliminate the possibility of intended/unintentional biases by physicians, patients, and/or statisticians. For example, in a RCT comparing two new therapies to a placebo, different tags (represented by shapes in Figure 1) may be associated with each treatment (i.e., \( a, b, \) or \( c \)). When assigning the treatments to a random sample of patients, only the tag (shape) is known, not the actual treatment (see Figure 1). A one-factor model is commonly used to analyze such a problem considering the treatments as fixed effects. If the treatments are randomly assigned to the tags, then the treatments are random effects, and the target for inference is the realized treatment difference. Xu and Stanek (2006) study this problem, adapting the ideas of finite population mixed models to the RCT setting.

The first step in formulating a finite population mixed model in the RCT setting is defining the potential observable outcome \( y_{sa} \) for subject \( s \), \( s = 1, 2, ..., N \) under treatment \( a \), \( a = 1, 2, ..., A \), and expressing the \( N \times A \) matrix of such outcomes as \( y = (y_{sa}) \). Xu and Stanek (2006) consider the simple setting where there is no subject by treatment interaction, so that \( y_{sa} = \mu + \beta_s + \tau_a \), with the row marginals of \( y \) given by \( y_s = \mu + \beta_s \), the column marginals given by \( \mu + \tau_a \), and \( \mu = \frac{1}{NA} \sum_{i=1}^{N} \sum_{a=1}^{A} y_{sa} \). Two sets of indicator random variables, \( U \) and \( V_{aA} = (v_{ja}) \) are introduced to express the permutation of subjects and the permutation of treatments, where \( v_{ja} \) takes on a value of one if treatment \( a \) is assigned to position \( j \) in a permutation, or zero otherwise. We use \( U \) and \( V \) to represent a joint permutation of units and treatments leading to \( Y = (\hat{Y}_s) = UyV' \). Then, we express

\[
Y = 1_N 1_A' \mu + 1_N T' + b1_A'
\]
where \( Y_i^* = \sum \sum U_{ia} V_{ja} y_{ia} \), \( T = V\tau \) (corresponding to random treatment effects), \( \tau' = (\tau_1, \tau_2, \ldots, \tau_A) \), and \( b \) is defined similar to (8). Defining \( Y^* = vec(Y') \), we re-write (26) as

\[
Y^* = X\mu + ZT + E
\]

where \( X = 1_{\times A} \), \( Z = I_a \otimes 1_N \) and \( E = (I_a \otimes I_N) b \). Using properties of permutation matrices, we obtain

\[
E_{y'}(Y^*) = I_{\times N} \mu \quad \text{and} \quad Var_{y'}(Y^*) = \sigma^2 J_A \otimes P_N + \sigma^2 A \otimes J_N \]

where \( \sigma^2 = \frac{1}{N-1} \sum_{i=1}^N \beta^2_i \), \( \sigma^2 = \frac{1}{A-1} \sum_{a=1}^A \tau_a^2 \), with \( \xi \) and \( V \) indexing expectation with respect to \( U \) and \( V \), respectively. As in the foundational representation of experimental designs by Zyskind (1962, 1967), random effects arise from permutations of finite populations, and can occur even when populations are as small as two, such as when \( A = 2 \).

Figure 1. Illustration of Tags (Square, Circle, and Triangle) Corresponding to the Assigned (but unobserved) Treatments and All Possible Assignments of Treatments in a Blinded Randomized Controlled Trial Comparing Two Treatments to a Placebo

We identify the sample as the random variables corresponding to \( A \) consecutive subsets of \( n \) PSUs from a permutation of subjects, where each subset is associated with a different tag. This corresponds to the random variables in (27) corresponding to consecutive sets of random variables in \( n \) rows (permuted subjects) shifted consecutively to different columns of \( Y \). Target quantities corresponding to linear combinations of \( T \) are given by \( P = g' Y^* \). The BLUP of \( P \) is obtained by re-arranging the random variables in \( Y^* \) into the sample and remaining portions similar to Section 3.22, resulting in the BLUP of the difference between the realized treatments assigned to tags \( j \) and \( j' \) as

\[
\hat{P}_{jj'} = k(\bar{Y}_j - \bar{Y}_{j'})
\]

where \( k = \frac{n\sigma^2 + \sigma^2}{\sigma^2} \) and \( \bar{Y}_j = \frac{1}{n} \sum_{i=1}^n y_{ij} \), the sample mean for \( T_j \). The predictor given by (28) has a smaller MSE (by a factor of \( k \)) than an estimator of the same difference based on a fixed effects model.

Treatments are considered to be random effects in (27), even though all treatments are included in the study, avoiding the unnecessary restriction of Kempthorne (1975) that effects be fixed when all are sampled. The predictor is identical to the predictor of Henderson, Kempthorne et al. (1959) as summarized by Henderson (1984) for the difference between two random effects, but the variance accounts for the finite population and finite number of treatments. Initial simulation studies developed by Xu (cluster web site, 2008) illustrate that using empirical estimates of \( k \) diminishes the advantage of (28) over the simple mean difference. Current research is exploring this area further, and expanding the experimental design model to include subject by treatment interaction as well as endogenous measurement error.
5. RESEARCH DESIGN AND METHODS

This research builds on over 10 years of collaborative NIH-funded research (R01-HD36848, Stanek, PI and R01-HL071828-02, Stanek, PI). The aim of our research continues to be the development of stochastic models that formally incorporate known information on the structure of the population under study and the mechanism of obtaining study samples in the context of public health sample surveys, observational studies, experimental clinical trials, and translational research. We use such models to develop optimal estimators/predictors of target quantities of the population under study. The ambitious research allows a wide variety of practical problems to be framed using this structure. The structure is fundamental for many biological and public health applications because it directly links the physical setting to the stochastic models used to develop inference via parameter definitions, facilitating interpretation of results. The methodological work is unified in the finite population mixed model discussed in the background, and varied in its application as described in the preliminary studies and progress reports.

5.1 Steps in Methodological Development

The proposed research will be developed using guidelines that have proved successful in past research, and will take advantage of the unique knowledge and experience of the study team in formalizing random variables and notation that incorporate sampling into the finite population model framework. The first step in addressing a new problem is clear identification of the problem, and an application drawn from NIH-funded clinical and translational studies at UMASS and elsewhere, providing real data examples. Next, we will define the problem in a non-stochastic setting using the simplest possible situation. This definition will identify hierarchical units if they exist, distinguish different study conditions, and include possible endogenous measurement error. Using the expected response, parameters and notation will be defined for the population. Derived parameters, as in (21), will be defined along with parameters that represent population variances. Simple assumptions will be used for the moments of measurement error; typically we will assume that the expected value is zero, and that they are uncorrelated. As problems increase in complexity, we will begin by making simplifying assumptions that retains the problem’s basic essence. The problem definition may include counterfactuals, with emphasis given to problems where counterfactuals are potentially observable, as in an experimental design setting.

Next, we define the process that gives rise to the potentially observed response. When subsets of units are observed, indicator random variables will be defined, as in (4), and will subsequently be used to identify random variables and units, retaining the remaining unobserved variables that define the population. The elementary random variables will be used to write a model for a single response as in (22), as well as a model for the joint response (23). We will investigate whether projections of the random variables to a smaller set of random variables enable us to obtain estimators/predictors of the target quantities without loss of information as in Stanek and Singer (2008). Using such random variables, BLUP will be developed of target quantities using methods given in Section 3.22. Simple examples will be constructed to illustrate the difference between the predictor and the competitors. Finally, we will study the properties of the new predictor relative to competitors and identify settings where substantial gains occur.

In the final step, we compare the performance of empirical predictors obtained by replacing the shrinkage factors by estimates relative to the performance of competitors via simulation studies, as well as construct interval estimators. Choices of parameters for simulations will be guided largely by probable ranges occurring in typical clinical and translational research project. Method of moment estimates of variance components will be used to construct the empirical predictors. We will complement the simulations with analytical approximations to the MSE of the empirical predictors via Taylor series expansions. The results will identify contexts where the new predictors generate the largest gains in efficiency (as in San Martino, Singer et al. 2008) as well as settings where empirical predictors result in loss of efficiency that is greater than the theoretical gains (as in Li, Stanek et al. 2008). To illustrate their use, we will apply the results to practical examples from the epidemiological literature and existing data sets (such as the Seasons Study and Physical Activity in Pregnancy Study, and the Step Ahead Study) for dietary, physical activity, and cardiovascular risk factors.

We plan to develop interval estimators for predictors of population means (or totals) using bootstrap and possibly jackknife variance estimation methods (Royall and Cumberland 1978; Wolter 1985; Särndal,
Swensson et al., 1992; Sitter 1992a,b). Since there is some empirical evidence that bootstrap methods track one tailed error rates better than jackknife methods (Li 2003), we will focus on them (Rust 1985; Kovar, Rao et al. 1988; Sitter 1992b; Rust and Rao 1996; Shao 1996; Tang 1999; Yeo, Mantel et al., 1999; Morris 2002), implementing the methods via software platforms such as SAS, Stata, and R, taking advantage of existing code. We plan to validate algorithms via preliminary simulations using the third grade population data for regression estimators used by (Valliant, Dorfman et al., 2000, p321 (www.wiley.com/public/sci_tech_med/finite_population/file: finite_pop.zip). We also plan to use the Seasons and Physical Activity in Pregnancy Study data sets in designing the simulations.

5.2 New Research Focuses

We plan new research that will build on past work (lead researcher, secondary researcher):

a) **Accounting for endogenous and exogenous measurement error** including the development of expanded finite population models and predictors for such settings (Singer, Stanek);

b) **Development of guidance for practitioners** to identify settings where increased (or decreased) accuracy will result from accounting for auxiliary variables and confounding (Li, Stanek);

c) **Extension of models and predictors** to settings where auxiliary variables include measurement error (Li, Lencina);

d) **Development of applications of auxiliary variables** suitable for use in predicting domain means and while accounting for other auxiliary variables as in synthetic estimation (Stanek, Lencina);

e) **Incorporation of cluster level covariables** to develop predictors for targets corresponding to realized adjusted cluster means, measures of association, or regression coefficients (Stanek, Singer);

f) **Longitudinal study settings**, including growth curve problems (Singer, Stanek);

g) **Extension to broader classes of experimental designs** including group randomized trials (Stanek, González);

h) **Investigation of applications with missing data** (Stanek, Li);

i) **Extension to non-linear models**, including the development of applications of two-stage cluster sampling settings for prediction of realized cluster means when response is categorical (Singer, San Martino)

Based on our past experience (see preliminary studies/progress), we anticipate that this process will result in new insights that will impact statistical practice, lead to clearer interpretation of statistical results, and foster new methods of nonparametric analysis applicable to a wide range of public health research.

We plan to accomplish this research using an experienced research team working over a five-year period. Each researcher will take the lead in a problem area, collaborating with at least one other team member. This focused approach will enable progress in multiple areas, with results shared at biannual working group meetings. The working group meetings will allow cross-fertilization of insights and strengthen individual research thrusts. The five-year time period will enable us to capitalize on the synergy, leading to new developments. Such a time period is necessary in light of the broad range of problems being investigated, and the inter-connectedness of the problems through the finite population mixed model method. The research effort will focus on the development of methods for practical use, and the exposition of those methods via manuscripts in the methodological and applied peer reviewed literature.

a) **Accounting for endogenous and exogenous measurement error including the development of expanded finite population models and predictors for such settings**

We plan to develop models that further our understanding of how account appropriately for endogenous and exogenous measurement error when predicting latent values or estimating target parameters. We also plan to provide guidance for practical applications where such error arises. When there is endogenous measurement error, subjects and measures constitute a hierarchy similar to that in clustered populations. Differences in exogenous measurement error are tied to sample positions, and relate to stratification. We will develop these ideas using new models that account for endogenous error as in (11) and exogenous error as in (13). The results will be applied and illustrated in the Season’s study where dietary intake is measured via 24-hour interviews by nutritionists.
In addition, we plan to develop expanded models that account for the two sources of measurement error simultaneously. Such models may be represented by random variables similar to (2) given by

\[ \begin{align*}
U_{11} (y_1 + W_{1r}) + \tilde{W}_{1r} & \quad U_{12} (y_2 + W_{2r}) + \tilde{W}_{2r} & \quad U_{13} (y_3 + W_{3r}) + \tilde{W}_{3r} \\
U_{21} (y_1 + W_{1r}) + \tilde{W}_{2r} & \quad U_{22} (y_2 + W_{2r}) + \tilde{W}_{2r} & \quad U_{23} (y_3 + W_{3r}) + \tilde{W}_{3r} \\
U_{31} (y_1 + W_{1r}) + \tilde{W}_{3r} & \quad U_{32} (y_2 + W_{2r}) + \tilde{W}_{3r} & \quad U_{33} (y_3 + W_{3r}) + \tilde{W}_{3r}
\end{align*} \]

Initial investigation reveals that more information is included in this larger set of \( N^2 \) random variables when used to predict a realized subject effect (see cluster web site, c07rz01.doc and c07rz04.doc). We will systematically apply a theorem due to Rao and Bellhouse (1978) to identify projections of these random variables that are sufficient for such purposes, characterize the properties of empirical predictors via simulation, and develop interval estimates for target quantities.

When the response error distribution is finite and discrete, the extension can be directly related to two-stage sampling, where the distribution of SSUs corresponds to the response error distribution. For example, suppose each subject selected via srs is asked to report the number of days in the past 7 day period that s/he smoked marijuana, with a single response reported for each subject. The number of days reported by the subject may differ from the actual number of days (latent value) that the subject smoked marijuana in the past 7 days; and the difference between these values is endogenous response error. Given that a subject’s response would range from 0 to 7 days, and the actual number of days smoked is an integer (also ranging from 0 to 7), the potential response error is a set of integer values, the distribution of which is not likely to be uniform. Considering all possible responses as a finite set, we can represent response errors by a random variable corresponding to a random selection from an expanded number of potential responses, and use methods for two-stage sampling to predict the latent value. For example, for a subject who actually smoked marijuana on 3 of the past 7 days, the potential response may correspond to the set of \( M=12 \) equally likely values, \((1,2,2,3,3,3,3,3,3,4,4,5)\), such that a response of 3 has a 50% chance of occurring. Extensions will be developed corresponding to with-replacement sampling of response, and multiple response.

Finally, we will extend results to situations where, by design, an unequal number of measurements of the response variable is planned for different selected PSUs. Such a situation occurs when a question is repeated for a subset of the sampled subjects as part of a longer interview, with a goal of assessing measurement error. A similar problem arises in a study where, by design, all sampled subjects should complete an equal number of interviews (i.e. \( m=3 \) 24-hour diet recalls), but some recalls are missing for some subjects. A simple approach for including multiple responses is to work with average responses, and redefine the variance component appropriately. Explicitly including unequal numbers of responses for subjects will require an expanded set of random variables when accounting for sampling.

b) Development of guidance for practitioners to identify settings where increased (or decreased) accuracy will result from accounting for auxiliary variables and confounding

Li, Stanek et al. (2008) used simulations to study properties of the empirical predictor of smoking prevalence given by (18) adjusting for a single binary variable (gender) concluding that the unadjusted estimate of prevalence is often more accurate. Guidance is needed for researchers to identify settings where control for confounding improves accuracy in situation where there are continuous covariables and mixtures of continuous and categorical covariables. We will develop such guidance using simulation studies, including advice on tradeoffs involved in categorizing continuous covariables.

c) Extension of models and predictors to settings where auxiliary variables include measurement error

Estimates of association have been developed by González, Singer et al. (2008) to account for endogenous measurement error which simplify to (18) when \( Q=1 \). When multiple auxiliary variables are available, there is often interest in estimating an adjusted measure of association which controls for confounding/balances additional covariables. Two settings can be distinguished: cases in which covariables are discrete and are used to classify subjects (such as gender or age), and cases where covariables are continuous, such as diastolic blood pressure. Such settings are valuable to distinguish given that the meaning of adjustment may
differ, and corresponding to combining stratum specific measures in one case, and corresponding to association of adjusted values in the other. We will develop predictors of an adjusted measure of association in these contexts, beginning with a categorical covariable.

The model used by González, Singer et al. (2008) for association, i.e., (19), included endogenous measurement error for the response, but assumed that auxiliary variables were measured without error. In many applications, such as when evaluating the association between LDL cholesterol and saturated fat intake, endogenous measurement error occurs for the independent variable in addition to response. Model (19) can be extended to include such error as

$$Z^* = X(\mu, 0') + I_N(b, d) + (E')_{N \times d}$$

where $W = (W_1, W_2, ..., W_d)$ and $W_d$ is an $N \times 1$ vector of response errors assuming $r = 1$ for each $s = 1, ..., N$. In such settings, we will develop estimators for association of the latent value of response with the latent value of a single auxiliary variable, and extend these methods to multiple auxiliary variables, resulting in estimators of adjusted measures of association, evaluating the results using simulation studies. Initial investigations with a single auxiliary variable with endogenous measurement error indicate that for an auxiliary variable with modest reliability (0.75) and a strong correlation with latent response (0.9), the variance of the predicted association may be inflated by nearly 2.5 fold due to the measurement error. Because many auxiliary variables are self-reported, the results warrant further investigation. Important applications occur when endogenous measurement errors are present in both response and auxiliary variables such as body mass index calculated based on self-reported body weight and height, and nutritional intake, as in applications using the BRFSS data.

d) Development of applications of auxiliary variables suitable for use in predicting domain means and an extension of such methods to settings accounting for other auxiliary variables for strata or domains as in synthetic estimation

Lencina, Singer et al., (2007) developed a conditional approach for domain estimation. We plan to develop an alternative approach avoids conditioning on the domain sample size, adapting Li’s 2003 approach for adjusting for auxiliary variables. Let $x_{hs}$ represent an indicator variable that has a value of one if subject $t^* = 1, ..., N_h$ in stratum $h$ is in domain $s = 1, ..., I$, and zero otherwise. Since $N_h = \sum_{t=1}^{N_h} N_{hs}$, the nested representation of subjects in domains, i.e. $s!^*$, has a one-to-one correspondence with the simple subject index in stratum $h$, i.e., $t$. When domain size is equal for all domains in a stratum, we use these indicator variables to define a response vector for subject $t$ in domain $s$ given by $y_{hs}^* = (x_{h1s}y_{hs}, x_{h2s}y_{hs}, ..., x_{hNs}y_{hs})'$, which we summarize for stratum $h$ by $Y_h^* = (y_{h1}^*, y_{h2}^*, ..., y_{hNs}^*)'$, a matrix that represents response for subject $t$ in domain $s$ in row $t$, column $s$, with all other elements equal to zero. A random permutation of subjects is defined by introducing the indicator random variables, resulting in $Y_h^* = U_hY_h^*$. Concatenating a column expansion of $Y_h^*$ for all strata results in a framework where domain means can be defined without conditioning on the sample domain sizes.

This approach may be applied to predict response in small areas, avoiding the need to use ratio estimates that consider domain sample sizes random, or condition on the domain sample size. Extensions of this approach to include additional covariables, as in Holt, Smith et al. (1979), will be developed by augmenting the response with additional auxiliary variables. Using this approach, we propose to extend predictors of domain means to settings where response is augmented by indicator stratum random variables. We will apply the results to estimate community-specific body mass index (continuous, $kg / m^2$) and obesity (binary) prevalence and compare the results obtained usual (hierarchical) mixed models. Because the finite population mixed model methods account for known finite population structure and actual sampling process in addition to characteristics of population composition that are used in conventional mixed models (see Li, Land, et al., 2008, Li, Kelsey et al., 2008), we anticipate that estimates based on finite mixed models should outperform those based on conventional models.
e) Incorporation of cluster level covariables to develop predictors for targets corresponding to realized adjusted cluster means, measures of association, or regression coefficients

We will extend the simple finite population setting that includes covariates to the two-stage cluster sampling problem, using the results on auxiliary variables and association described in Preliminary Studies. First, we will consider cluster level covariates considering, for example, a setting where clusters of patients correspond to physicians' practices. The cluster-level covariates may include the physicians' specialties, gender, race, years of practice, affiliation to a medical school, languages spoken, etc. and may be categorical (gender, race, specialty) or continuous (age, years of practice). Other examples include studies of Medicare and Medicaid and HMO reimbursement policy (such as pay-for-performance physician incentive plan), where interest may focus on estimating a primary care physician (PCP)’s patient satisfaction rating based on a small sample survey of the PCP’s patients in a year, or estimating how patient satisfaction rating is associated with a PCP’s characteristics; or estimating compliance rate by auditing a random sample of medical charts. Accuracy of the estimates has important financial implications to both practicing physicians and HMOs (American College of Physicians 2007, National Committee for Quality Assurance 2007).

To incorporate a single cluster level covariate (i.e. $Q=1$), we will extend (15) with the substitution of $U$ by

$$\left(U \otimes I_M\right)\left(\bigoplus_{s=1}^N U(s)\right)$$

and $(y \ x_s)$ by variables that represent an equal size clustered population where

$$y = \left(y_1 \ y_2 \ \cdots \ y_N\right)', \ y_s = \left(y_{s1} \ y_{s2} \ \cdots \ y_{sm}\right)', \text{ and } x_s = \left(\delta_1 \ \delta_2 \ \cdots \ \delta_N\right) \otimes I_M'$$

where $\delta_s = x_s - \mu$, for $s = 1, ..., N$, resulting in the random variables

$$\left(U \otimes I_M\right)\left(\bigoplus_{s=1}^N U(s)\right)\left(y \ x_s\right).$$

We will use a column expansion of these variables to represent response in the population, define the target as a (possibly weighted) linear combination of these variables, divide these variables into sample and remainder, and use the prediction theory as in Section 3.22 to develop the optimal predictor of the target parameter. Subsequent work will evaluate the predictor relative to competitors via simulation studies, and develop interval estimates of the predictor.

When only cluster-level covariates are included in the model, there will be no variation for the covariate within each cluster and the derivation is anticipated to be significantly simpler. Extending the definition of $x$ to accommodate cases where its elements vary with subjects and clusters, we will extend the development to measures of association with characteristics of SSUs. Applications include estimating hospital-level average patient functional improvement returns after total knee replacement surgery based on a national registry (Zimmer TKR Registry) of TKR patients accounting for hospital characteristics, and adjusting such estimates to account for surgery variables, such as a surgeon’s experience (Franklin, Li et al., 2008).

Predictors of PSU means have been developed by Stanek and Singer (2008), but involve complicated functions of the sample and population values. Further work is needed to identify practical approaches that can be implemented. Several possibilities will be investigated. The challenge is the development of empirical estimates for shrinkage constants that include functions of PSU means. Strategies to be considered include reparameterizing the shrinkage constants in terms of coefficients of variation, and developing alternative predictors that are less efficient but feasible to implement in practice. In addition, iterative predictors will be considered.

f) Longitudinal study settings, including growth curve problems

A common problem occurs when clusters correspond to subjects, and repeated measures correspond to consecutive time intervals for a subject. In this context, mixed models developed under an 'infinite' population of subjects and times are usually considered for analysis. Finite population mixed models are important in this context since subject-specific results are commonly of interest over a given (finite) time period (such as one year).
For example, in longitudinal studies, a cluster may correspond to a group of days (as in a week, month, or trimester of pregnancy), with the target parameter corresponding to the average response for a subject (such as nutrient intake, physical activity, or marijuana use) over the defined time period (Stanek, Well et al. 1999). In such studies, the target parameter may be causally linked to other variables of interest (such as weight, cholesterol, or future use of cocaine/heroin). The target parameter itself may be defined for a cluster of measures or time interval, motivated by biological or behavioral theoretical models. The value of a subject’s parameter is often a subject’s risk factor. Examples include a subject’s saturated fat intake relative to serum cholesterol (in the Season’s study NIH/NHLBI R01-HL52745 PI: Ockene, Merriam, Ockene, et al., 1999; Ockene, Chiriboga, et al., 2004); a subject’s exercise relative to weight gain in a pregnancy trimester (Schmidt, Erickson, et al., 2002) (NIH/NIDDK R01-DK074876 PI: Chasan-Taber); or changes in LDL cholesterol subsequent to a systems-based nutrition and counseling program in the Watch Study. In worksite studies, the research interest may focus on estimation of prevalence of absenteeism or work productivity in relation to obesity, diet quality and physical activity (e.g., nurses working in an Orthopedic Department).

We will develop the finite population mixed model for longitudinal studies, beginning with the simplest context, where a population is composed of clusters, and a cluster mean is defined as the average of the expected response at two fixed times. For example, clusters may correspond to patients, and response may be measures of fasting total serum cholesterol at both baseline, and one-year follow-up. Estimates of the change in cholesterol for sample subjects, and estimates of the average change in the population are of interest. Such examples occur frequently in health-related research.

We will consider two strategies for including time in the model: One strategy postulates that the model for a cluster (i.e. subject) includes a parameter at each time point, in addition to response error. This is a multivariate approach common in repeated measures studies including growth curves (Pothoff and Roy 1964, Grizzle and Allen 1969, Rao 1987, von Rosen 1995, Morrell, Peterson et al., 1997). A second strategy considers time as a SSU level covariate in the model. This approach extends analyses accounting for covariates outlined in Section 5.2g.

We describe the first strategy when response is reported for a sample of subjects at two time points, with primary interest in the expected gain in response. The approach is closely related to Sections 4.2 and 4.4, where the two variates for each subject correspond to gain and initial response. Suppose the population average initial (baseline) response is known, with each subject’s initial response measured without error. Endogenous measurement error is present at the second time point, and included in the gain in response. For each subject, consider the gain in response as the outcome, and the initial response as the auxiliary variable. The optimal predictor of gain in the population can be obtained by taking weights equal to one using (20). Predictors of gain for subjects can be obtained by setting \(c = c_e\) (see Section 4.5) when defining the target. This approach will be generalized to \(Q\) time points, and extended to settings where the auxiliary baseline variable is measured with error, to settings where measurement times differ between subjects, to growth curve problems (where average higher order trends are zero), and to finite population random coefficient models.

We sketch the proposed second strategy, using as illustration prediction of the change in weight over a one-year time interval in free-living heathly men and women (Chiriboga, Ma, et al., 2008). We first represent weight for subject \(s\) at time \(t_M = t\) for \(t = 1, \ldots, M\) as \(y_{st}\), or equivalently by \(y_{st} = \beta_{1s} + \beta_{2s}t_s + \epsilon_{st}\). The parameter \(M\) defines the potentially observable set of responses, and the interpretation of \(t\). We define subject parameters corresponding to average subject’s weight and average time of measurement as \(\mu_s = \sum_{i=1}^{M} y_{st} / M\) and \(\tau_s = \sum_{i=1}^{M} t_{st} / M\) (noting that in this special setting, \(\tau_s = 6.5\) for all \(s = 1, \ldots, N\) ). The change in weight over one year based on the subject specific regression model is defined by \(\beta_{1s} = \sum_{i=1}^{M} w_{st}y_{st}\) where \(w_{st} = (t_{st} - \tau_s) / \sum_{i=1}^{M} (t_{st} - \tau_s)^2\).

Notice that \(\beta_{1s}\) is the subject-specific population total of \(z_{st} = w_{st}y_{st}\) for \(t = 1, \ldots, M\) as in Section 4.4. Assume that a simple random sample of \(n\) subjects will be selected, with four weight measures scheduled to be made on each sample subject (as in the Seasons Study). The measurement times for a sample subject may arise
from different study designs (e.g. corresponding to a systematic sample, a simple random sample, a
deterministic set of points, an experimental plan, etc). For simplicity, we assume that time points are selected
via simple random sampling for each subject, resulting in a two-stage cluster sample with weight measured at
$m = 4$ times for each sample subject.

We can predict the annual change in weight, or the change in weight for a realized sample subject by replacing
the response variable in Section 4.5 by $z_{it}$, and using the two-stage cluster sample prediction approach.
Additional research will expand the second stage sampling design beyond simple random sampling, and
account for additional subject specific covariables and time dependent covariables, building on results of
Sections 4.2, 4.5, and 4.6.

g) Extensions to broader classes of experimental designs

We will begin by extending results to factorial design with subject by treatment interactions and endogenous
measurement error, building on previous results. Next, we will account for additional covariables, either with
known or unknown population means. We will also develop methods for multi-group problems with stratified
populations, to pretest-posttest randomized controlled trials and to clustered randomized trials possibly
including a repeated measures designs. As time permits, extensions in each of these settings will include
additional covariables and additional longitudinal applications, closely relating results to Section 5.2.f.

h) Investigation of applications with missing data

In the simplest context, suppose the same finite set of possible response errors (that sum to zero) is
associated with each subject in a finite population. Let us represent the missing data mechanism via an
independent Bernoulli random variable $X_{it}$ that takes on a value of 0 (if missing) or 1 for each potential
response. Using this idea, data for response $t$ for subject $s$ can be viewed as a bivariate pair, $y_{it}X_{it}, X_{it}$,
where $X_{it}$ follows a Bernoulli distribution. The sample data are generated by a two-stage process: in the first
stage, subjects are sampled, while in the second stage, the bivariate pairs are selected. Such a process
underlies data that are missing at random, and include the intentional missing data process (due to sampling)
and the un-intentional missing data process (due to non-response). This general framework can be expanded
to include non-response mechanisms that depend on various subject characteristics, some of which may be
used in analysis. A prediction-based approach will be used to estimate linear functions of the corresponding
random variables.

Other frameworks that represent the missing data process may be conceived. We will investigate such
frameworks and develop basic results in the simplest settings. As such efforts succeed, extensions will be
made to settings where the Bernoulli parameter depends in some fashion on $y_{it}, H_s$, characteristics of the
subject, or factors related to the sample selection.

i) Extensions to categorical and non-linear models

We will explore strategies for adapting the finite population mixed model to settings where response is
categorical or where the cluster mean is a multiplicative function of the population mean. Settings where
response is categorical will be compared with non-linear model alternatives when predicting cluster means.
These applications can make direct use of the nonparametric finite population sampling model results.
Non-linear models will begin with a review of the literature for multiplicative models applied to the finite population
context. This review will include the GSK approach (Grizzle, Starmer et al., 1969) approach implemented in
SAS Catmod, as well as the estimating function approach outlined by Binder and Patak (1994). Work in this
area with application to finite populations distinguishes ‘cluster-specific’ models from ‘population-averaged’
models, with distinct cluster parameters for the two models. The relationship between cluster parameters in
the two models has been approximated by (Neuhaus, Kalbfleisch et al., 1991) and discussed by Holt, Smith et
al. (1989). We will identify possible design-based approaches based on the finite population mixed model that
enable non-linear models, such as the logistic regression model, to be fit to two-stage clustered sample data.
This aspect of the research will initially be exploratory, and first integrate non-linear model approaches with the finite population mixed model.

5.3. Time line and responsibilities

Structure for the Research Operation

Our past work has provided us valuable experience that will guide the development of future work. First, this research will make use of the combined expertise of a research team with a track record of successful collaboration at the University of Massachusetts (UMASS), and the University of Sao Paulo (USP), Brazil. Former doctoral students at UMASS and USP will have an active role in the research (including Dr. Li at UMASS, Dr. Lencina at Universidad de Tucuman and Dr. San Martino at Universidad Mar del Plata in Argentina), as will new doctoral students at each institution. Dr. Stanek at UMASS, and Dr. Singer in Brazil will coordinate the activities. This collaboration has been facilitated by development and maintenance of an extensive research WEB site, frequent communication via Email, and by Skype conference calls.

Most importantly, the research has been fostered and enhanced by biannual face-to-face two-week working meetings (January at USP and July at UMASS). All researchers will attend each working meeting, along with doctoral students at the home institution, and invited experts with interest and expertise in the research areas. At the UMASS meetings, Drs. Bertone-Johnson, Chasan-Taber, and Ockene will participate and provide input and context for applications. At each research meeting, a set of research papers will be prepared in a bound workbook and distributed to participants. At the most recent meeting in Natal, Brazil, a total of 36 working papers and manuscripts were included in the workbook (including over 500 pages). One half of each day in the first week of the meetings is devoted to formal presentations of current research by investigators, usually with three presentations per day. Discussion of the research by the group leads to revisions, extensions and additional applications. Local and external experts are invited to give presentations and provide feedback. The meeting focuses on development of methods and applications for publication in the peer-reviewed literature.

The research WEB site (www-unix.oit.umass.edu/~cluster/) will be maintained at UMASS Amherst, and will catalogue developing manuscripts and reports, identify current directions and ideas, and document interactions. The WEB site contains an electronic version of all workbook documents, providing broad access for other investigators. The WEB site helps promulgate research results presented at national and international meetings and in peer-reviewed journals. We also plan to develop a portion of the site for applications, and to develop material to facilitate access for other researchers. The WEB site will be promoted at presentations national and international meetings.

The research agenda is ambitious. We plan to pursue extensions to finite population mixed model methods in a deliberate, staged fashion. We have outlined the scope of the research, realizing that methodological developments may be easier in some areas, and require more in depth work in other areas. The lead and secondary researcher are identified with areas previously given. We will constantly re-evaluate the direction of methods development and application to areas that appear feasible and have the potential to have a practical impact. At the conclusion of each biannual working group meetings, we will re-focus the research objectives. Research objectives with a primary and secondary collaborator will be set for the subsequent six month period. Dr. Stanek will direct the project and work closely with Dr. Singer who will serve as project director at USP in Brazil. Dr. Stanek will also be responsible for supervising the maintenance and enhancement of the project WEB site, and logistic support for the Amherst working group meetings. Dr. Singer will be responsible for logistic support at the USP working group meetings.