Introduction to Revision of "Mixed Models for Finite Populations"

This is the second revision of this proposal. The original proposal, entitled "Improving estimation in clustered randomized trials" was unscored. Based on comments of the review, and continuing work by the investigating team, a completely revised version of the proposal, entitled "Mixed Models for Finite Populations" was submitted, and received a score of 2.24.

This revision of the proposal responds to the concerns expressed by reviewers in the second review, and augments the background, preliminary results, and research plan based on new work by the investigators. In particular, in response to some of the reviewers concerns, and to reviews of a manuscript contain the main finite population mixed model results submitted to JASA, the finite population mixed model methods are more concretely situated relative to mixed model and super-population model literature. We have developed theoretical expressions that quantify the reduction in expected mean squared error (MSE) of the finite population mixed model predictors relative to competing predictors, and to characterize settings where this reduction will be large (i.e. greater than 50%). As a result of these efforts, the manuscript (Stanek and Singer, 2004) describing these results has been accepted for publication in JASA. These new results directly compare finite population mixed model predictors with other predictors, and illustrate their strengths.

Additional research that we have conducted since the second review further addresses concerns of the reviewers. To evaluate the impact of mixed model methods in practice, we have developed a modular simulation program (similar to the simulation study conducted by Pfeffermann and Nathan (1981)) that enables empirical versions of the finite population mixed model predictor to be compared against the empirical mixed model BLUP and the empirical version of Scott and Smith's super-population model predictor. Initial results of the simulation indicate that the superior properties of the finite population mixed model predictors are retained when shrinkage constants are estimated from the sample. Although additional study is needed, these preliminary results make concrete a direct comparison of the methods which was not available in the earlier proposal.

Several master and doctoral students are currently engaged in research related to this proposal. A recently completed thesis (Lu, 2004) developed methods of incorporating missing data into the random permutation model framework. This innovative work develops a strategy for explicitly accounting for missing data mechanisms that operate on fixed units, as distinct from missing data mechanisms that operate on 'selected units', or random effects. The proposed research will enable such work to be more rapidly developed and extended.

Several changes have occurred to the study team which will strengthen the proposed research. Dr. Li has joined Dr. Reed at UMASS-Worcester, simplifying collaboration between investigators at the UMASS-Amherst and UMASS Worcester-campus (separated by 50 miles). In addition, collaborators in Brazil will all be located at a single institution, the Department of Statistics at the University of Sao Paulo, headed by the co-investigator, Dr. Singer. As suggested by the reviewers, we have made changes in the staffing that have resulted in reductions in the budget (such as the elimination of a data manager and masters level RA positions), while still retaining an adequate level of PhD and faculty support to accomplish the methodological work.

The research team has continued to pursue and develop our research agenda since the last submission. One year ago, Dr. Stanek traveled to the University of Sao Paulo for one week to work with Drs Singer and Bolfarine. This past January, Dr. Singer visited the Biostatistics and Epidemiology program at UMASS-Amherst to work with Drs. Stanek, Li, and Reed for two weeks. This pattern of close collaboration based on periodic face to face working visits will be continued as part of the proposed research plan. The travel costs to support these meetings are modest (~5%). We continue to involve content specialists (Epidemiologists and a Cardiologist) in our research, and plan to draw on five primary research studies for applications.

We include response to the individual reviewer's comments next, and follow this response with a summary of changes that have been made in the proposed research.

Responses to Concerns Raised by Reviewer 1.

The reviewer recognized the broader focus of the re-submission, namely mixed models for finite population, with cluster-randomized trials being just one of the application areas. We retain this broader focus, and have developed a more specific research plan for each of the areas. We note that although the focus of this research is on biomedical applications of mixed model methods, the methodology is directly applicable to survey sampling, and modeling in complex surveys. While not altering the research focus, we will develop manuscripts as part of this research that will promote this cross-fertilization.

We plan to develop finite population mixed models to repeated measures and longitudinal settings, and have clarified a framework for the extensions by defining clusters, units, and conditions (which may correspond to time (Section D.1)). Since the number of potential applications is broad, we have defined the manner in which we plan to extend the results, using applications to motivate the extensions. To provide a fuller plan for these extensions, we have re-organized Section D to outline the extensions first (in Sections D1-D5), and then presented the strategy we plan to use to attack the problems (in Section D6).

The reviewer noted that a “significant amount of progress has been made in pushing ahead” the research since the original grant submission. We feel that we have made an additional significant amount of progress in pushing ahead our research since the last submission. As part of this progress, we are able to more concretely motivate and compare the prediction methods, a concern raised by the reviewer.

We have motivated application of the methods in Section B.1, where we illustrate a practical setting (bullying) where predictor of random effects in a finite population is needed (Figure 1). We have directly compared the theoretical expected MSE between the predictors in Section B.2 (Figure 2) illustrating that the size of the improvement that can be expected using a finite population mixed model predictor may be large. A more extensive set of comparisons is given in the manuscript by Stanek and Singer (2004 in press). An indication of the theoretical developments that give rise to the comparisons is given in Section C.2.2.

We have added description of many concrete settings where finite population mixed models would apply drawing on research studies described in Section E. These settings are used to help motivate and situate the need for methodological development in Section D. They also will provide a source of applications. We agree with the reviewer’s assessment that the available data from research studies are quite interesting.

We have dropped from the proposed work development of bootstrap and jackknife methods for interval estimation, focusing instead our attention on developing estimates of the expected mean squared error of an empirical predictor, and the coverage of an interval estimator based on such an estimator. We have changed our focus due to reviewer’s concerns, and since we believe the approach will provide a more straightforward way of quantifying uncertainty in the predictor. We plan to evaluate the coverage properties of interval estimates using such an approach in simulation studies.

We have responded to concerns by the reviewer over the budget. We have eliminated the data manager and one masters level RA from the project, as suggested by the reviewer. The project personnel have also been reduced. This has resulted in an overall reduction in the budget.

Responses to Concerns Raised by Reviewer 2.

The second reviewer felt that there was insufficient support to the claim that there is a problem in using standard the mixed model methods, and that “The impact may be great..”. We agree with the reviewer that the previous submission did not document clearly the impact of using a finite population mixed model predictor relative to a standard mixed model predictor, or a super-population model predictor. Additional methodological work that we have conducted since the previous submission has allowed us to quantify the magnitude of the gain (in terms of reductions in expected MSE) that will occur using the finite population mixed model predictor. We illustrate this gain in one special case in Figure 2 (section B.2), and provide the general expression that can be used to make the theoretical comparisons between predictors in Section C.2.2. Additional comparisons of the theoretical reduction in expected MSE using the finite population mixed model predictor are presented by Stanek and Singer (2004, in press in the Appendix). We feel that this additional work and the results address the reviewer's concern over the lack of evidence to support our claim, and strengthens the proposal.

A second concern by the reviewer was a concern that we did not "well-motivate…(the proposal) …using statistical issues arising in biomedical studies…". We have amplified on examples that motivate the proposed research. For example, Figure 2 (section B.1) illustrates a practical study where interest is naturally focused on the change in a realized random effect in a finite population context. The research that we propose to develop will be directly relevant to answering the question as to magnitude of change in bullying in the realized school. We have re-organized the research and methods section to motivate each of the areas of methodological research with a biomedical application. We note that the research team has the benefit of a rich set of research studies available from which such applications can be selected.

A third concern expressed by the reviewer was that the research would not produce software that would enable others to use the methods. We have directly addressed this concern, and plan to develop macros that will enable others to readily implement the methods. We anticipate that developing such macros will be relatively easy, since the predictors are linear functions of the sample data, and empirical estimates of variance components are available from other standard procedures. We discuss the issues and strategy for such development in Section D.6.4. Such macros will be publicly available on the Project WEB site, and highlighted in publications and presentations. Thus, we feel that the methods have the potential to affect statistical practice.

The reviewer expressed concern that there was inadequate attention to the literature on auxiliary variables when developing methodological strategies. We acknowledge that a good knowledge of the literature is critical. An important paper in this regard is by Pfeffermann and Nathan (1981). We have critically reviewed this paper (along with others) and are incorporating ideas from such work in our research. It is noteworthy that none of the previous works attempt to build models in a design-based framework. This highlights the innovativeness of the proposed research. Due to space limitations, it was not possible to provided a detailed review of such work, but a more complete summary is given by Stanek and Singer (2004 in the Appendix).

The reviewer noted that the mixed model, super-population model, and finite population mixed model may not actually estimate the same quantities. We agree with this assessment, and feel that this concern provides a strong rationale for using the finite population mixed model methods. With such methods, the quantities that one is interested in estimating are defined in a finite population. For this reason, the quantities have clear interpretations and may be of practical relevance. A similar argument does not apply to the abstract quantities that may be the target of inference using other methods.

The reviewer felt the review and discussion of resampling, jackknife and bootstrap estimators was limited. We agree with this assessment, and have dropped this from the proposed research. We will focus instead on developing estimates of the expected mean squared error of an empirical predictor, and the coverage of an interval estimator based on such an estimator. We have changed our focus due to reviewer's concerns, and since we believe the approach will provide a more straightforward way of quantifying uncertainty in the predictor. We plan to evaluate the coverage properties of interval estimates using such an approach in simulation studies.

Note: Changes in the proposal are indicated in *italics*.

Mixed Models for Finite Populations

A. Specific Aims

Mixed models are widely used to analyze data from observational, experimental, and longitudinal studies where subjects in the underlying population define different clusters (such as communities, clinics, physician practices, schools, classrooms, or families). *Covariate associated with the cluster or the subject (e.g. gender, age) are important to include in such analyses. Subjects may be measured under different conditions, as in longitudinal studies, where repeated measures (e.g. days in a month) can be considered a cluster of occasions, with different conditions (time) associated with individual occasions (days).*

Frequently, the number of clusters defining the study population is finite and cluster sizes are finite and unequal. Current mixed model methods (Crowder and Hand 1990; Murray 1998; Brown and Prescott 1999; McCulloch and Searle 2001; Diggle, Heagerty, Liang and Zeger 2002; Raudenbush and Bryk 2002) do not account for the finite population and cluster sizes. While ignoring the finite population context, such methods may often be applied in unequal size clustered sample settings. *This research will develop new statistical methods for mixed models that are appropriate for in such finite population contexts, accounting for the finite population structure.*

Models that account for the finite population and cluster sizes have been proposed using a super-population framework in survey sampling (Scott and Smith 1969; Prasad and Rao 1999), but are not widely used in modeling clustered populations, in part due to the artificial nature of the super-population *and lack of software*. Finite population mixed models that do not require a super-population framework have been recently developed (Stanek and Singer 2004). *These predictors are non-parametric and out perform mixed model and super-population model predictors, in some cases dramatically so.* Their foundation is the two-stage random permutation underlying a two-stage sample design, and can be applied to continuous and discrete data. ***The aim of this research is to evaluate and extend statistical methods for the finite population mixed model. The research is innovative in developing non-parameteric design-based methods for estimation and prediction in clustered studies. It is important since the predictors have smaller MSE, require minimal assumptions, and provide results specific to the studied population.***

The finite population mixed models we consider arise from explicit representation of random variables underlying two-stage random permutations of the population. Previous research has developed such models to estimate population parameters and predict random effects in two-stage samples from populations of equal size clusters with response error (Stanek and Singer 2004). *Additional work has developed* predictors of random effects in populations with unequal size clusters without response error (Stanek and Singer 2003). This research will:

- *extend the finite population mixed model to include balanced and unbalanced replicated response error, and to include response error in populations with unequal cluster sizes.*
- *develop finite population mixed model methods to include cluster/unit level auxiliary variables to improve predictors of realized random effects and to develop estimators of regression parameters for the auxiliary variables.*
- *develop and extend finite population mixed model methods to longitudinal settings, and settings where units are observed under more than one condition.*

As part of this research, we will *connect theory for expansion and projection of random variables that distinguishes labeled clusters from random selections of clusters, with a focus on* unequal sized clustered populations, unifying principals for finite population mixed model developments. We will also evaluate the improvements (increased precision) with respect to other proposed methods, evaluate the impact of replacing shrinkage factors with estimates, develop *estimates of the expected mean squared error (MSE) of the predictors, and evaluate the coverage of interval estimates based on the estimated MSE.*

The research will lay the groundwork for key extensions of the methodology to longitudinal settings, *inclusion of cluster and unit level* covariates, and expanded random variable frameworks. The research will provide a foundation for future extensions to experimental designed studies such as cluster randomized studies, transition models that include autoregressive response, as well as extensions to unequal probability designs.

B. Background and Significance

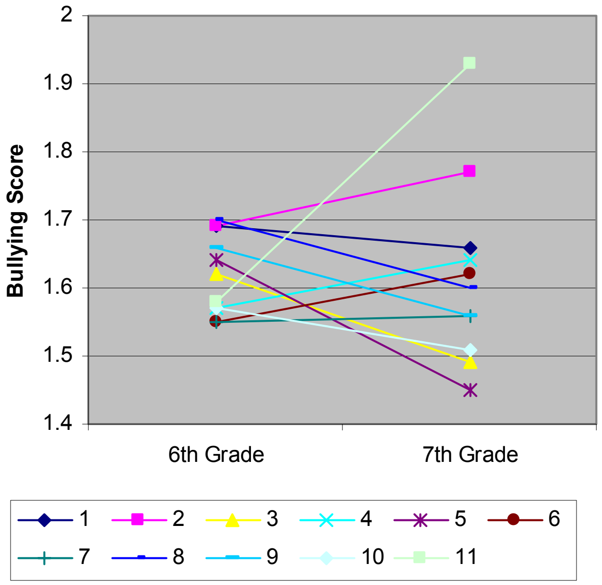
B.1. Introduction

Mixed models, characterized by inclusion of fixed and random effects, are used in many contexts in public health. Fixed effects correspond to population parameters. Random effects are used to represent parameters for clusters of similar units (such as communities, work sites, schools or classrooms, families, or clinic patient groups), or subjects (where units correspond to repeated measures). Cluster parameters may correspond to the expected cluster response, or specific regression parameters (as in a random coefficient model). Mixed models may be applied when there are equal and unequal numbers of observations per cluster to estimate fixed effects and predict linear combinations of fixed and random effects. However, current methods for mixed model analysis do not account for the finite population structure, since the models are formulated in an infinite population setting.

There are many settings where finite population clustered data arise, but the finite structure is not incorporated in the analysis. Natural examples occur where cluster sizes are small, and measures are made on an appreciable fraction of the cluster members. Clusters made up of children in families, physician practices in hospitals, classrooms in schools, and carpal tunnels in subjects are obvious examples. Examples include studies with clustered study designs such as the Watch study (Ockene, et al. 1996) *to evaluate the impact of a nutritional and exercise intervention on cholesterol in Worcester, Ma, or the Urban School Norms Study, where a classroom intervention in a random sample of middle schools seeks to change student norms for behavior and substance use (Schensul,* R01-DA12015).

There is often interest in estimating response for individual clusters. For example, as part of the Urban School Norms study, the development of negative behavior was studied longitudinally in a sample of 11 middle schools among students in 6th grad (in 2000-2001) and 7th grade (in 2001-2002). Each student was asked to score nine questions such as "During this school year, other kids in school called me names or swore at me," as 1=not at all; 2=once; 3=2-3 times; or 4=4+ times. An average of the nine responses formed an index of bullying for each student each year. The number of students in a school varied from 63 to 328, and the fraction of students measured per school varied from 26% to 63%. Figure 1 summarizes the average bullying scores in 6th and 7th grade by school. In the study, in keeping with the random assignment of schools to intervention or control conditions, individual schools are considered to be random effects for analysis. Although the main focus of the study is on the effectiveness of the intervention, there is a natural interest in school specific changes in bullying behavior. Parents, teachers, and principals would be interested in knowing which school is School #11 (displaying the largest increase in bullying scores in Figure 1), and whether the change is statistically significant. The increase in average bullying score (+0.35) for this school is diminished to a predicted increase of +0.07 using standard mixed model methods and the best linear unbiased predictor(BLUP). However, this predictor does not account for the different sampling fractions of students in schools, or the finite (and different) sizes of the schools themselves. Methods that account for such differences will be developed in this research.

Figure 1. Average Bullying Scores by School



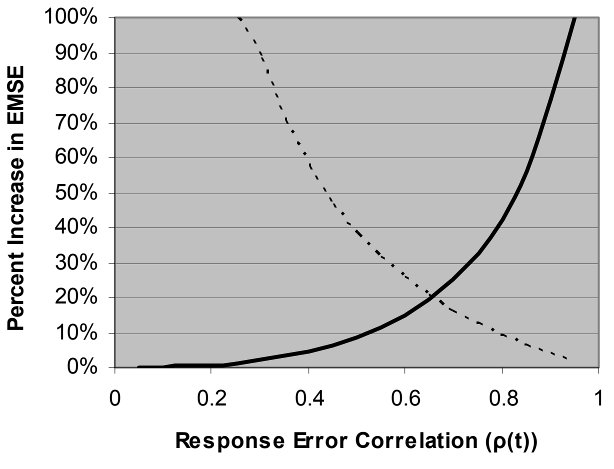
There are other examples where careful problem definition reveals a clustered finite structure. In longitudinal studies, a cluster may correspond to a group of days (as in a week, month, or trimester of pregnancy), with the target parameter corresponding to the average response for a subject (such as nutrient intake, physical activity, or marijuana use) over the defined time period. 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), by a biological or behavioral theoretical model specifying the time interval that defines the cluster. The value of the parameter for a subject is often a subject's risk factor. Examples include a subject's saturated fat intake relative to serum cholesterol (in the Season's study (Merriam, et al. 1999; Ockene, et al. 2004); a subject's exercise relative to weight gain in a pregnancy trimester (Schmidt, et al. 2002) (Physical Activity in Pregnancy, R03-HD-393441, Chasan-Taber PI); and a subject's marijuana use relative to subsequent cocaine or heroin use (Schensul, et al. 2000)(the Pathway study, R01-DS11421,Schensul PI). Such parameter definitions are typically not formulated as finite population parameters, in part since statistical analysis methods are not specific to different formulations. This research will evaluate and extend methods that can be tailored to account for parameters defined over different time periods, and hence address this methodological gap.

B.2. Problems with Predictors from Traditional Mixed Model and Super-Population Model

Although mixed models are widely used, usual methods are not tailored to finite populations. Common applications of mixed models often assume normality for the distribution of clusters, and units within clusters, and postulate a simple response error model for a unit. These assumptions imply an infinite size population. Scott and Smith (1969), using a model based super-population approach, have developed methods that appear to account for a finite size population. However, when compared with predictors developed from a finite population mixed model (Stanek and Singer 2004), both the standard mixed model and Scott and Smith's predictors may perform poorly.

The three predictors of a realized cluster latent value were recently compared by Stanek and Singer (2004) in terms of the expected MSE. The finite population mixed model predictor is always best, where the magnitude of the reduction in MSE depends on the fraction of units selected in a cluster, f, the intra-class correlation, $\rho(s)$*, and the response error correlation* $\rho(t)$*. As illustrated in Figure 2 (when the population contains 100 cluster, each consisting of 20 units, with 30 clusters selected in a sample) the expected MSE (EMSE) of the BLUP of a realized cluster mean using mixed model methods can be over 150% the EMSE of the random permutation (RP) predictor (i.e. based on finite population mixed model methods) (Stanek and Singer, 2004). The EMSE is also elevated for Scott and Smith's predictor relative to the finite population mixed model predictor. These theoretical results indicate that substantial improvements can be made to mixed model and super-population model methods.*

Figure 2. Percent Increase in EMSE for Mixed Model (——) and Scott and Smith Model (- - -) Relative to the RP Model Predictors: [f=0.9 ρ(s)=0.1]



In practice, predictors of random effects require estimates of shrinkage constants, resulting in empirical predictors. There is evidence that the superior theoretical performance of finite population mixed model predictors is retained in practice based on a simulation program developed for two-stage sampling. For example, using empirical predictors with f=0.6, $\rho(s)=0.33$*, and* $\rho(t)=0.2$*, the simulated (based on 1000 trials) EMSE for the RP model empirical predictor was 0.96, compared with an EMSE of 1.21 for the empirical mixed model, and an EMSE of 0.99 for the empirical Scott and Smith predictor (Stanek, Singer, Li, and Reed, 2004). This represents 26% and 3% reductions in the EMSE compared to the mixed model, and Scott and Smith model, respectively.*

B.3. Traditional Mixed Models

Mixed models are widely used to analyze data arising from clustered populations or from clustered study designs. The mixed model methods were developed from work by (Eisenhart 1947; Kempthorne 1955; Scheffe 1956a; Scheffe 1956b; Scheffe 1959) and expanded and extended by (Henderson, Kempthorne, Searle and von Krosigk 1959; Dempster, Laird and Rubin 1977; Harville 1977; Harville 1978; Laird and Ware 1982). The methods have been popularized by software (such as SAS Proc Mixed (Littell and Wolfinger 1995, and MLwiN (Goldstein 2003)), and more recently by numerous texts (Crowder and Hand 1990; Brown and Prescott 1999; Verbeke and Molenberghs 2000; McCulloch and Searle 2001; Diggle, Heagerty et al. 2002; Raudenbush and Bryk 2002). The methods enable traditional estimation of fixed parameters, as well as prediction of random effects using best linear unbiased predictors (BLUPs) as discussed by (Goldberger 1962; Henderson 1984; McLean, Sanders and Stroup 1991; Robinson 1991; Stanek, Well and Ockene 1999; McCulloch and Searle 2001). The sample data are typically considered to have been selected from a conceptual infinite population, often vaguely defined, but understood to be the limit (as the size becomes infinite) of a very large finite population of interest.

B.4. Frameworks for Modeling Finite Populations

Traditional methods for inference for finite population parameters have been based on sampling designs (Konjin 1962; Cochran 1977; Hansen, Madow and Tepping 1983; Hansen, Dalenius and Tepping 1985), and often focus on unbiased estimation of the population mean or total (Horvitz and Thompson 1952). Following Godambe's (1955) results, other approaches have been developed. Super-population model based methods were proposed, such as (Hartley and Rao 1968; Hartley and Rao 1969; Hartley and Sielken 1975), or (Royall 1968; Royall 1970; Royall and Herson 1973a; Royall and Herson 1973b). Generalized regression estimator (GREG) (Cassel, Sarndal and Wretman 1976; Sarndal, Swensson et al. 1992) and calibration estimators (Sarndal and Wright 1984; Deville and Sarndal 1992; Brewer 1999) were also developed. The model assisted approach has not resulted in a unified framework for inference (Brewer, Hanif and Tam 1988; Skinner 1989; Brewer 1995; Brewer 1999). In contrast, the super-population model based approach of (Royall 1973; Royall 1976a; Royall 1976b; Royall 1986; Royall 1988) has been developed into a unified approach for inference (Valliant, Dorfman and Royall 2000). Advocates of both approaches have recognized limitations of their approach, and attempted to improve them by adopting aspects of the other approach (Brewer 1979; Sarndal and Wright 1984; Royall 1992; Brewer 2002). Related to this work are the model based approaches for nested settings developed by (Fuller and Battese 1973; Bellhouse, Thompson and Godambe 1977; Pfeffermann and Nathan 1981; Pfeffermann 1984; Bellhouse 1987; Stukel and Rao 1997; Bellhouse and Rao 2002).

B.5. Mixed Models for Finite Populations

B.5.1. Bayesian and Super-population Model Approaches

Although there was some early discussion of mixed models in finite population settings (Henderson 1984)(Chapter 15), mixed model methods have been extended principally using a Bayesian paradigm in a hierarchical setting, or via super-population models. Hierarchical Bayes estimation has been developed for finite populations following work by (Ericson 1969; Ghosh and Meeden 1986; Datta and Ghosh 1991; Datta and Ghosh 1995) and summarized by (Ghosh and Meeden 1997)(Chapter 5). More recently, a Bayesian approach using an operational likelihood for finite population parameters has been proposed by (Bolfarine, Gasco and Iglesias 2003).

The Bayesian approach can be distinguished from the super-population model approach. Both approaches focus on prediction, and ignore design probabilities once the data have been observed. The super-population model approach uses frequency based methods to predict parameters in a finite population (defined as a realization from a super-population). It does not require specification of a prior distribution.

The starting point for the super-population model approach is the super-population (of random variables) from which the finite population is a realization (Royall 1976; Bolfarine and Zacks 1992; Valliant, Dorfman et al.

2000). Upon sampling, a subset of random variables is realized (the sample), while the remaining random variables are not observed. Predictors (linear functions of the sample) of linear functions of the un-observed random variables are developed under the restriction of unbiasedness and minimum mean squared error. Such an approach was used by (Scott and Smith 1969) for two-stage sampling, extended to include covariates by Pfeffermann and Nathan (1981) and extended to include response error by (Bolfarine and Zacks 1992). The methods have had limited application outside of survey sampling, in part due to their sensitivity to model miss-specification, and the artificial nature of the postulated super-population.

B.5.2. Design Based Mixed Models for Finite Populations

Bayesian and super-population model methods do not account for design probabilities and require postulating prior distributions, or an artificial super-population. An alternative approach is a finite population mixed model that arises directly from random variables linked to an identified finite population and sample design (Stanek and Singer 2004). In two-stage sampling, the sample corresponds to a partial realization of a two-stage random permutation of the population. We refer to models based on this probability framework as **finite population mixed models**. Since the models account for the finite population and design *but require no additional assumptions, they are non-parametric. The linkage between target parameters in the population and predictors is maintained in the theoretical development.*

Such methods have been recently developed using two-stage random permutation models for equal size clustered population with response error (Stanek and Singer 2004), and unequal size clustered populations (Stanek and Singer 2003). Essentially, the development uses the sampling design to link the population to a set of random variables representing permutations of the units in the population. The random variables arising from the design play the role of random variables in a ‘super-population’. One difference is that these random variables are based on the sampling design, and hence the model is design based. The variance structure is also different, with variance parameters defined in the finite population.

Representing a permutation as a vector of random variables, Royall’s (1976) approach is used to develop estimators of population parameters and predictors of random variables. The inferential framework is not limited by the need to specify an artificial super-population or specify prior distributions. The methods require minimal assumptions (simple random sampling) and directly account for the finite population structure.

B.6. Significance of this research.

This research will impact inference from finite population studies where mixed models are applied. The impact can be traced to conceptual differences between the finite population mixed model, and alternative models, as indicated in Table 1 in the Preliminary Results. Using finite population mixed model methods (unlike usual mixed models) only un-observed units are predicted. Unlike super-population model methods, random variables in the finite population mixed model *enable response error to be separated from lack of fit, and accounted for explicitly in the predictors.* This provides for a direct links between predictors and finite population parameters without additional assumptions.

This research will extend finite population mixed model methods by developing new theoretical predictors, by evaluating these predictors (and similar empirical predictors) via simulation studies, and by developing software macros that can implement these predictors using SAS and STATA. We will extend models for unequal size clustered populations to include response error, and multiple (and unequal) numbers of responses. We will extend models to longitudinal settings, including random coefficient models. Extensions will also be made to practical contexts where cluster level covariates are present, and where unit level covariates occur.

Empirical predictors and simulation studies will build on the modular simulation program developed by Stanek (2004) for balanced two stage cluster sampling (see section C.6). We will draw on five research studies (see Section E) for applications and context (eg., dietary and physical activity assessments). Simulation studies will evaluate finite population mixed model empirical predictors relative to the usual mixed model and super-

population model predictors. The studies will also evaluate the impact of variance component estimation on interval estimates for predictors.

This research will be conducted by an established research group with a track record of collaboration and publications in this area, providing a high potential for success. Preliminary results and a clear research plan are evident for each proposed extension of the methods. New developments will begin with the simplest settings and be expanded in a disciplined manner, insuring feasibility and an achievable project scope. The range of extension, along with the evaluation of the finite population mixed model predictors relative to those resulting from usual normal theory mixed models, and two-stage super-population models, will provide a broad context basis for applications. Finally, this work will provide a foundation for extensions to experimental designed studies, extensions to longitudinal transition models, *non-linear models*, and extensions to unequal probability sampling designs.

C. Preliminary Studies/Progress Report

C.1. Introduction

The background for this research was developed as part of a NIH funded study entitled "Improving Analysis Methods for Cluster Randomized Prevention Trials" from 9/97-7/02 (R01-HD36848, Stanek, PI, and including Drs Singer and Bolfarine). We summarize significant results of this research that form the foundation for this proposal.

C.2. Basic notation and the finite population mixed model.

We first describe the basic terminology and notation that we will use for a finite clustered population with response error. We assume that the number of units in each cluster is equal, and that the observed data corresponds to a two-stage simple random clustered sample. Of interest are parameters that represent averages or totals in the population, or averages or totals of units in individual clusters.

We define a finite population by a listing of *M* units, indexed by $t = 1,...,M$ in each of *N* clusters, indexed by $s = 1,...,N$, where the k^{th} response for unit t in cluster s is given by $Y_{stk} = y_{st} + W_{stk}$. The model for unit t in cluster s is a response error model where y_{st} is a fixed constant representing the expected response for the unit, and W_{stk} represents response error (with zero expected value). We parameterize response for unit t in cluster s as

$$y_{st} = \mu + \beta_s + \varepsilon_{st} \quad (1)$$

where $\beta_s = \mu_s - \mu$ is the deviation of the cluster mean, $\mu_s = \frac{1}{M}\sum_{t=1}^{M} y_{st}$, from the population mean, $\mu = \frac{1}{N}\sum_{s=1}^{N} \mu_s$, and $\varepsilon_s = y_{st} - \mu_s$ is the deviation of the expected response for unit t in cluster s from the cluster mean. This parameterization is called a derived model (Hinkelmann and Kempthorne 1994). Combining these models, $Y_{stk} = \mu + \beta_s + \varepsilon_{st} + W_{stk}$. Note that only W_{stk} is a random variable in this model.

We assume that a two-stage cluster sample is to be selected from the population (without replacement), with m units selected in each of n clusters. On each selected unit, we assume that a single measure is to be made of some response. The sample is a partial realization of a two-stage random permutation of units in the population.

C.2.1. The finite population mixed model

We represent the population under a two-stage random permutation model as an ordered list of NM random variables, where both clusters and units in clusters are independently permuted. For each permutation, we assign a new label, $i=1,...,N$ to the clusters according to their position in the permuted list. In a similar manner, we label the positions in the permutation of units in a cluster by $j=1,...,M$. Without loss of generality, we define the sample as the first m positions for units in each of the first n positions for clusters in a permutation. Using this notation, a model for a unit in the j^{th} position in a cluster in the i^{th} position is represented by $Y_{ijk}^{*}=Y_{ij}+W_{ijk}^{*}$. For different permutations, the particular cluster and subject in the ij^{th} position will vary. If we randomly select a permutation, the cluster at the i^{th} position in the permutation of clusters and the unit at the ij^{th} position are, in this sense, random. For ease of exposition, we refer to the cluster that will occupy position i in the permutation of clusters as primary sampling unit i (PSU), and to the unit that will occupy position j in the permutation of units within a cluster as secondary sampling unit j (SSU). PSUs and SSUs are indexed by positions (i and j), whereas clusters and units are indexed by labels (s and t) in the finite population. Since any unit in any cluster may occupy position ij , we represent the expected response (over response error, ξ_3) for SSU j in PSU i as the random variable $E_{\xi_3}\left(Y_{ijk}^{*}\right)=Y_{ij}$. Using the parameterization in (1), we define the **finite population mixed model** as

$$Y_{ijk}^{*}=\mu+B_i+E_{ij}+W_{ijk}^{*} \quad (2).$$

Prior to sample selection (which can be viewed as selecting a permutation), the cluster that will correspond to PSU i is random, and hence the expected response for the cluster in the i^{th} position is a random variable. Once the sample (i.e. permutation) has been selected, it will be apparent which cluster corresponds to a particular PSU. We refer to that cluster as the **realized PSU**, and the average of the expected response for the units in that cluster as the **latent value** for the realized PSU. Thus, the latent value for the realized PSU i is the realized value of $\mu+B_i$.

The finite population mixed model is formalized by introducing indicator random variables that relate y_{st} to Y_{ij}. The indicator random variable U_{is} takes on a value of one when cluster s is assigned to PSU i, and a value of zero otherwise; the indicator random variable $U_{jt}^{(s)}$ takes on a value of one when unit t in cluster s is assigned to SSU j in PSU i, and zero otherwise. As a consequence, the random variable corresponding to SSU j in PSU i in a permutation is given by $Y_{ij}=\sum_{s=1}^{N}\sum_{j=1}^{M}U_{is}U_{jt}^{(s)}y_{st}$, while $B_i=\sum_{s=1}^{N}U_{is}\beta_s$ is a random effect that represents the deviation of the latent value for PSU i from the population mean. *Representing* one measure *for* each SSU, using vector and matrix notation, the mixed model for the finite population is given by

$$\mathbf{Y}^{*}=\mathbf{X}\mu+\mathbf{ZB}+\left(\mathbf{E}+\mathbf{W}^{*}\right)$$

where $\mathbf{X}=\mathbf{1}_{NM}$ and $\mathbf{Z}=\mathbf{I}_N\otimes\mathbf{1}_M$ where $\mathbf{1}_a$ is an $a\times 1$ column vector of ones and $\otimes$ denotes the Kronecker product (Graybill 1983). This model represents random variables for the entire population, not simply the portion of the population that is part of the sample. Using properties of the indicator random variables (denoting expectation over PSUs and SSUs by ξ_1 and ξ_2, respectively), $E_{\xi_1\xi_2\xi_3}\left(\mathbf{Y}^{*}\right)=\mathbf{X}\mu$ and $\text{var}_{\xi_1\xi_2\xi_3}\left(\mathbf{Y}^{*}\right)=\left(\sigma_r^2+\sigma_e^2\right)\mathbf{I}_{NM}+\sigma^{*2}\left(\mathbf{I}_N\otimes\mathbf{J}_M\right)-\frac{\sigma^2}{N}\mathbf{J}_{NM}$, where $\mathbf{J}_a=\mathbf{1}_a\mathbf{1}_a'$, σ_r^2 is the average response error over all SSUs, σ_e^2 is the average variance of the expected response of SSUs within PSUs, σ^2 is the PSU variance, and $\sigma^{*2} = \sigma^2 - \frac{\sigma_e^2}{M}$.

C.2.2. Prediction in the finite population mixed model.

Linear combinations of the expected response for SSUs, $T = \mathbf{g}'\mathbf{Y}$, define population parameters (such as the population mean when $\mathbf{g}' = \mathbf{1}'_{NM}$), and PSU latent values (such as the *expected value* of PSU i when $\mathbf{g}' = \frac{1}{M}\mathbf{e}_i{}' \otimes \mathbf{1}_M{}'$, where $\mathbf{e}_i$ is an N-dimensional column vector with a value of one in the i^{th} position, and zero elsewhere). Stanek and Singer (2004) use the prediction approach of (Royall 1976) to develop estimates of parameters and predictors of random variables. The predictors are best (have minimum prediction error), linear (in the sample), and unbiased, i.e. BLUP. The BLUP are developed by first partitioning the random variables $\mathbf{Y}^*$ into a sampled and remaining set so that T can be expressed as the sum of a sample and remaining term. When response error is not present, since the sample will be observed (without error), attention is limited to predicting the remaining term. When response error is present, both the sample and remaining terms need to be predicted. The predictor is determined by minimizing the prediction error, subject to an unbiased linear constraint. The BLUP can be derived for any linear combination of $\mathbf{Y}^*$ or $\mathbf{Y}$.

Of particular interest, for example, is a comparison of BLUP for the latent value of PSU using the finite population mixed model with predictors developed from other approaches (Table 1).

Table 1. Predictors of the Latent Value of PSU i when $i \le n$ in Two-stage Cluster Sampling

Model	**Predictor**
Mixed Model	$\hat{P} = \left(\hat{\mu} + k_i\left(\bar{Y}_i - \hat{\mu}\right)\right)$
Scott&Smith	$\hat{P}_i = f\bar{Y}_i + \left(1-f\right)\left(\hat{\mu}^* + k_i\left(\bar{Y}_i - \hat{\mu}^*\right)\right)$
Finite Mixed Model	$\hat{T}_i = f\bar{Y}_i + \left(1-f\right)\left(\bar{Y} + k\left(\bar{Y}_i - \bar{Y}\right)\right)$
Finite MM+Resp.Err.	$\hat{T}_i = f\left(\bar{Y} + k_r^*\left(\bar{Y}_i - \bar{Y}\right)\right) + \left(1-f\right)\left(\bar{Y} + k^*\left(\bar{Y}_i - \bar{Y}\right)\right)$

Additional terms in Table 1 correspond to the proportion of sample SSUs, $f = \frac{m}{M}$, sample means defined by $\bar{Y}_i = \sum_{s=1}^{N} U_{is}\bar{Y}_s$ where $\bar{Y}_s = \frac{1}{m}\sum_{j=1}^{m} Y_{sj}$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i$, weighted means, $\hat{\mu}$ and $\hat{\mu}^*$, where weights depend on the model assumptions. In addition, each predictor includes a different shrinkage constant, where the constants are functions of the variance components given by σ_r^2, σ_e^2, σ^2, and σ^{*2} *given by* $k_i = \frac{m\sigma^2}{\sigma_i^2 + m\sigma^2}$, $k = \frac{m\sigma^{*2}}{\sigma_e^2 + m\sigma^{*2}}$, $k_r^* = \frac{m\sigma^{*2} + \sigma_e^2}{m\sigma^{*2} + \sigma_e^2 + \sigma_r^2}$, *and* $k^* = \frac{m\sigma^{*2}}{m\sigma^{*2} + \left(\sigma_e^2 + \sigma_r^2\right)}$.

Comparing predictors between the finite population mixed model, and the usual mixed model, we see that the traditional BLUP places all the weight on predicting the random variables associated with the unobserved SSUs. Ignoring the proportion of the SSUs that are observed has the disadvantage of increased model sensitivity due to reliance solely on predictors based on the model. For example, when there is no response error, if a high fraction of SSUs, say 90%, are observed for a selected PSU, a large portion of the units comprising the PSU are known, and need not be predicted. Mixed model approaches act as if all units must be predicted.

The predictor developed by (Scott and Smith 1969) is similar to the finite population mixed model predictor, but differs in the shrinkage constant. Under a super-population model with equal SSU variances for all PSUs, if M is very large, $k \cong k_i$ and $\bar{Y} = \hat{\mu}^*$. In other cases, the predictors differ.

A larger distinction between predictors occurs when response error is present. For the usual mixed model and Scott and Smith's model, adding response error simply changes the shrinkage constant, not the form of the predictor. In the finite population mixed model, a shrinkage factor is evident for the sampled SSUs, in addition to a different shrinkage for the SSUs not realized. Since sampled SSUs are not known without error, this form of a predictor matches the intuition that some shrinkage should occur.

Since all predictors in Table 1 are linear functions of the sample, the expected MSE can be expressed as a quadratic function of c given by

$$MSE\left(\hat{T}\right)=\left(1-f\rho_t\right)\left[\frac{\sigma_e^2}{nm\rho_t}+\left(\frac{n-1}{n}\right)\left(1-k^*\right)\sigma^2\right]+\left(\frac{n-1}{n}\right)\frac{\sigma^{*2}}{k^*}\left(c-\left[f\rho_t+\left(1-f\rho_t\right)k^*\right]\right)^2$$

where $f=\frac{m}{M}$, $\rho_t=\frac{\sigma_e^2}{\sigma_e^2+\sigma_r^2}$ *and values of c are given in Table 2.*

Table 2. Values of c for Predictors $\hat{T}=\bar{Y}^*+c\left(\bar{Y}_i^*-\bar{Y}^*\right)$ of the Latent Value of PSU i when $i \le n$ in Two-stage Cluster Sampling with Homogeneous Unit and Response Error Variances.

Model	
Mixed Model	$c_{MM} = k_i$
Scott & Smith	$c_{SS} = f + \left(1-f\right)k_i$
Random Permutation.	$c_{RP} = f + \left(1-f\right)k$
Random Permutation with Response Error	$c_{RPR} = f\rho_t + \left(1-f\rho_t\right)k^*$

As illustrated in Figure 2 (Section B.2), a comparison of the expected MSE using these expressions reveals large reductions in the expected MSE in some settings (see Stanek and Singer 2004 for more discussion).

C.3. Expansion of Random Variables to Account for Unequal Size Clusters

Additional problems arise in representing a finite population mixed model with unequal size clusters. The problems stem from ambiguity of notation for ordered positions in a random permutation of SSUs relative to the underlying units and clusters. Since the dimension of a vector of cluster random variables depends on the cluster size, and these sizes differ for different clusters, when cluster vectors are permuted, the interpretation of which cluster corresponds to the j^{th} SSU in the i^{th} PSU in a vector representing a two-stage permutation of the population is problematic. We illustrate this with a simple example, and assume no response error. As in section C.2., we represent the fixed constant representing response for unit t in cluster s as y_{st}. Similarly, we represent the random variable corresponding to the i^{th} PSU and the j^{th} SSU in a two-stage permutation by $Y_{ij}=\sum_{s=1}^{N}\sum_{t=1}^{M_s}U_{is}U_{jt}^{(s)}y_{st}$. The distinction between the setting described in section C.2. is the dependence of the number of units in a cluster, M_s, on the specific cluster, cluster s.

For example, suppose that the population consists of three clusters such that $N=3$, $M_1=M_2=2$, and

$M_3 = 3$. Random permutations will permute clusters, and units in the clusters. The finite population, as well as examples of five possible realizations of a two stage permutation of the population are given in Table 3.

Table 3. Finite Population of 3 Clusters with $M_1 = 2$, $M_2 = 2$ and $M_3 = 3$, and Five Possible two-stage Random Permutations

Pop	$\mathbf{y}'$	y_{11}	y_{12}	y_{21}	y_{22}	$\boxed{y_{31}}$	$\boxed{y_{32}}$	$\boxed{y_{33}}$
$Perm$	1	y_{22}	y_{21}	y_{12}	y_{11}	$\boxed{y_{32}}$	$\boxed{y_{31}}$	$\boxed{y_{33}}$
$Perm$	2	y_{21}	y_{22}	y_{12}	y_{11}	$\boxed{y_{31}}$	$\boxed{y_{32}}$	$\boxed{y_{33}}$
$Perm$	3	$\boxed{y_{32}}$	$\boxed{y_{33}}$	$\boxed{y_{31}}$	y_{22}	y_{21}	y_{11}	y_{12}
$Perm$	4	$\boxed{y_{31}}$	$\boxed{y_{33}}$	$\boxed{y_{32}}$	y_{12}	y_{11}	y_{22}	y_{21}
$Perm$	5	y_{11}	y_{12}	$\boxed{y_{33}}$	$\boxed{y_{31}}$	$\boxed{y_{32}}$	y_{22}	y_{21}
	$\mathbf{Y}$	Y_{11}	Y_{12}	$Y_{??}$	$Y_{2?}$	$Y_{??}$	$Y_{3?}$	$Y_{3?}$

In the first random permutation (Perm=1), PSU 1 corresponds to the cluster $s = 2$, PSU 2 corresponds to the cluster $s = 1$, and PSU 3 corresponds to the cluster $s = 3$. For this permutation (Perm=1), the SSUs corresponded to the realized PSUs according to the original partitioning of the population. The same ordering of PSUs occurred for Perm=2.

For Perm=3, PSU 1 corresponds to the cluster $s = 3$, PSU 2 corresponds to the cluster $s = 2$, and PSU 3 corresponds to the cluster $s = 1$. With this realization, the third SSU in the population correspond to the first PSU, not the second PSU as in the first two permutations. This difference implies that SSU membership can not be unambiguously established by simply recording the order of the PSU and SSU, since the size of the PSU realized in a permutation determines the membership of SSUs.

The ambiguity caused by unequal size clusters in representing PSU membership in two-stage permutations can be avoided by expanding the representation of the random variables for a permutation. The expansion retains specification of the unit and the position for the random variables. We illustrate the basic idea of the expansion in the context of simple random sampling (as in (Stanek, Singer, and Lencina 2004) where random variables retain the identifiably of both label and position.

First, note that we can represent a random permutation of a finite population via a random permutation matrix composed of indicator random variables, such that $\mathbf{Y}^* = \mathbf{U}\mathbf{y}$ where $\mathbf{y} = \left(y_1, \cdots, y_N\right)'$ represents parameters in the labeled population, and $\mathbf{Y}^* = (Y_1, \ldots, Y_N)'$ represents a random permutation of the population, and $\mathbf{U} = \left(\left(U_{is}\right)\right)$ represents an $N \times N$ permutation matrix of indicator random variables. The expanded model is formed by setting $\mathbf{Y} = \left(\mathbf{D}_{\mathbf{y}} \otimes \mathbf{I}_N\right) vec\left(\mathbf{U}\right)$, where $\mathbf{D}_{\mathbf{y}}$ is a diagonal matrix with the elements of $\mathbf{y}$ along the main diagonal, $vec\left(\mathbf{U}\right)$ is a vector representing the column expansion of $\mathbf{U}$. The expanded random permutation vector $\mathbf{Y} = \left(\left(Y_{is}\right)\right)$ has dimension $N^2 \times 1$, with elements given by $Y_{is} = U_{is} y_s$. Note that the elements of $\mathbf{Y}$ include indices for both label and position.

In a simple random sampling setting, such an expansion enables linear combinations of random variables to be defined that represent the sum of random variables at a position, as well as the sum of random variables that correspond to a unit. Since these sums are different, they lead to interesting questions regarding interpretation of linear predictors (defined in terms of positions, as they usually are in exchangeable models). This is discussed more extensively in (Stanek, Singer and Lencina 2004).

C.4. Finite Population Mixed Models with Unequal Cluster Sizes

In the context of two-stage sampling, the expansion enables both PSUs and cluster labels to be identifiable in the resulting expanded set of random variables. This is the approach used by Stanek and Singer (2003 see appendix) to develop finite population mixed model predictors of PSU totals and means. We refer to details presented there, and summarize the resulting predictors of the total (or mean) for PSU i in Table 4. The first row in Table 4 gives the finite population mixed model predictor when cluster sizes are equal (for reference). In Table 4, $\bar{T}$ denotes the sample average of PSU totals and $\bar{W}$ denotes a sample average of PSU weighted totals. The constant c is arbitrary.

Table 4. Predictors of the Total or Mean of PSU i when $i \le n$ in Two-stage Cluster Sampling under Finite Population Mixed Models with Unequal Cluster Sizes

Parameter	**Predictor**
Mean (equal size)	$\hat{T}_i = f\bar{Y}_i \quad +(1-f)\left(\bar{Y} + k\left(\bar{Y}_i - \bar{Y}\right)\right)$
Total (unequal, pps)	$\hat{P}_i^{\circ} = T_i \quad +\left(\frac{1-f}{f}\right)\left(\bar{T} + k^{\circ}\left(T_i - \bar{T}\right)\right)$
Mean (unequal, pps)	$\hat{P}_i^{*} = W_i \quad +\left(\frac{1-f}{f}\right)\left(\bar{W} + k^{+}\left(W_i - \bar{W}\right)\right)$
Mean (unequal, not-pps)	$\hat{P}_i^{\bullet} = c\bar{Y}_i \quad +\left((1-c)\bar{Y} + ck^{\bullet}\left(\bar{Y}_i - \bar{Y}\right)\right)$

Shrinkage constants differ for different predictors, and depend on variance components corresponding to cluster totals, or weighted cluster totals (see Stanek and Singer 2003 in the Appendix for details). The difference in shrinkage constants can be attributed to the different variability that is evident if samples are characterized by means, totals, or other weighted values. The difference in predictors implies that if interest is focused on a PSU mean, a different predictor will result if the total for the PSU is predicted, and then divided by the number of SSUs in the PSU, as compared with predicting the PSU mean directly. These results have similar implications for prediction of population totals, or means. The target quantity that is to be predicted determines the definition of the random variables, resulting in different variance structures and hence different predictors. Using the finite population mixed model, methods can be more closely tailored to target parameters for which inference is desired.

C.5. Application of Seemingly Unrelated Regression Models to Account for Covariates

Li (2003) has recently developed a method for incorporating auxiliary information in a design based random permutation model setting. The development was made in the context of simple random sampling, but the general strategy has potential to be developed to other settings, most notably to two-stage cluster sampling. The results make use of a seemingly unrelated regression model for the response and the covariate. These main results are given by Li and Stanek (2004 see Appendix). The basic idea is to use a vec expansion of the random permutation model, $\mathbf{Y}_{N\times(p+1)} = \mathbf{U}\mathbf{y}$ where values of all $p+1$ variables for all N subjects can be represented with an $N\times(p+1)$ non-stochastic matrix $\mathbf{y}$, such that $\mathbf{y}_{N\times(p+1)} = \left(\mathbf{y}^{(0)} \mid \mathbf{y}^{(1)} \mid \cdots \mid \mathbf{y}^{(k)} \mid \cdots \quad \mathbf{y}^{(p)}\right)$.

The finite population prediction based approach is used to model a linear function of $\mathbf{y}^{(\mathbf{0})}$ based on the $k=1,2,\ldots,p$ auxiliary variables. Li (2003) focused on estimating the population mean or total. When there is a single auxiliary variable and the total is known, estimates of the population total correspond to regression estimators, i.e. $\hat{T}^{(0)} = N\left(\bar{Y}^{(0)} + \hat{\beta}_{01}\left(\mu^{(1)} - \bar{Y}^{(1)}\right)\right)$, where $\bar{Y}^{(0)}$ and $\bar{Y}^{(1)}$ represent the sample mean for the response and the co-variable, respectively, $\hat{\beta}_{01} = \hat{\sigma}_{01}/\hat{\sigma}_1^2$ if σ_1^2 is unknown or $\hat{\beta}_{01} = \hat{\sigma}_{01}/\sigma_1^2$ if σ_1^2 is known. Unlike survey sampling regression estimators, no linear regression model is required. More details and extensions are given by Li 2003.

C.6. Simulation Studies

Simulation studies have been developed to evaluate predictors of the expected response of a PSU based on a two stage random permutation model with response error, and used to evaluate the expected MSE of predictors and empirical predictors. The simulation is patterned after the simulation studies conducted by Pfeffermann and Nathan (1981). The simulation is written in SAS, and is based on modules corresponding to SAS macros. There are modules that generate a finite clustered population (with equal size clusters). The distribution of unit parameters can be patterned after different distributions (normal, uniform, beta, and gamma), with the standard deviation of the unit parameters allowed to be proportional to the cluster mean. Similar control is possible over the distribution of cluster means. Response error can be added to the units. Once the population is set, other modules select a two stage sample and evaluate predictors and their expected MSE.

The simulation study is documented in a number of technical reports (available in the reports for 2003 at http://www.umass.edu/cluster/), with the documentation currently being converted so as to be WEB based. Included in the development are strategies for implementing empirical predictors. This area of research is active, with new results continually being developed. The current understanding of these results will be presented at the March, 2004 ENAR meetings (Stanek, Singer, Li, and Reed 2004).

C.7. Missing Data

Recent work by Lu, Stanek, and Puleo (2004) has developed methods for explicitly accounting for missing data in a simple random permutation model. This work was part of a thesis by Lu (2004). A design-based prediction approach is used develop an estimator of the finite population mean in a simple setting where some responses are missing. The approach is based on indicator sampling random variables that operate on labeled units (subjects). The missing data mechanism may depend on the subject (such as may result from screen calls with an answering machine), or on a selection (such as when the sample is assigned to different interviewers). The estimator (which equals the sample total divided by the expected sample size) is developed as a predictor of the un-observed subjects using an approach usually reserved for model-based inference. The approach has a direct link to best linear unbiased predictors (BLUP) in finite population mixed models. When the probability of missing is estimated from the sample, the empirical estimator simplifies to the mean of the realized random variables. The different missing data mechanisms are revealed by the notation that accounts for the labels and sample selections. This work was limited to exploring setting where data are missing completely at random. The framework established, however, provides an avenue for extending results to contexts where more complicated missing mechanisms occur.

D. Research Design and Methods

D.1. Introduction and Overview

This research will develop and extend design-based finite population models, evaluate estimators and predictors obtained under such models relative to competitors, and apply such methods to cluster randomized trials and longitudinal studies. The methods are innovative since they provide a general framework for optimal prediction while accounting for finite population characteristics using a non-parametric sampling based probability model. The research will build on results that unify estimation and prediction in simple random sampling (Stanek, Singer, and Lencina 2004), on results that extend estimators of the total or mean to simple random samples with co variables (Li, Stanek 2004), on results used to develop predictors for primary *sampling unit (PSU) means and latent values in balanced two-stage cluster sampling with response error (Stanek and Singer 2004), and on and results that extend predictors PSU latent values in unbalanced two stage cluster sampling (Stanek and Singer 2003). The extensions will consider situations where:*

a. Response error, possibly with multiple observations per unit, is included in the model.
b. Auxiliary variables are used to improve the predictors or define regression target parameters.
c. Clusters and/or units (i.e., subjects) are observed under more than one condition.

This research will develop new predictors(estimators) for samples from a finite population, compare properties of the predictors relative to competitors from the literature, develop empirical predictors along with software for their implementation, and evaluate the performance of empirical predictors via simulation studies. The research will develop analytic methods that are optimal in simple random sampling, balanced two stage cluster sampling, and unequal size cluster sampling of finite populations.

Simple random sampling models will be extended to include response error, simple regression models with unit covariates, and repeated measures models that include lack of fit. Such models will be developed for pretest-posttest settings, and applied to observational studies where an intervening variable is non-randomly assigned to units. Models will also be developed for longitudinal settings corresponding to profile models and growth curve/random coefficient models. These developments are significant since (unlike other methods) they will distinguish sampling variability and lack of fit from response error.

Two-stage cluster sampling models will be developed that account for cluster specific covariates and unit-specific covariates in a balanced setting. We first will consider cluster specific covariates, extending settings to contexts where a cluster is observed under different conditions. These extensions pave the way for developing repeated measures and longitudinal analyses in a multi-stage cluster sample. Similar extensions are planned for unit-specific covariates.

Estimators/predictors for two-stage cluster sampling models with unequal size clusters will be studied in more depth by focusing on the synthesis of principles underlying expansion and projection of random variables, as described in C.3. This study will draw upon results from the related literature concerning ancillary variables, and orthogonal projections of nuisance parameters and apply them to the finite population mixed model setting. These results will be extended to include auxiliary variables and multiple conditions.

For clarity, we briefly summarize the general notation to be used in the proposed research. We study models for a finite population of units that may be further classified in clusters, where

s	*a specific cluster,* $s=1,...,N$
t	*a specific unit in a cluster,* $t=1,...,M_s$
c	*conditions associated with the unit and cluster, where* $c=1,...,C$ *(such as attributes of the cluster, unit,or situation). We assume here that a cross-classification of all levels of all covariables will form* C *conditions.*
u	*a co-variable,* $u=1,...,q$
k	*index of a measure of response,* $k=1,...,p_{st}$ *(response error)*
$\mathbb{N}=\sum_{s=1}^{N} M_s$	*Total number of units in the population*

In general, we represent the k^{th} *response on unit* t *in cluster* s *under condition* c *as* Y_{stck}. *The basic model is given by*

$$Y_{stck}=\mu_{stc}+W_{stck}\text{, where } E_{\xi_3}\left(Y_{stck}\right)=\mu_{stc}.$$

We may use a reduced parameterization where $\mu_{stc}=\sum_{u=1}^{q} x_{stu}\beta_u+\varepsilon_{stc}$, *with* ε_{stc} *corresponding to lack of fit.*

Two stage random sampling (see detail in C.2) is introduced into the model via indicator random variables U_{is} *(which takes on a value of one when cluster* s *will occupy position i in a permutation of clusters) and* $U_{jt}^{(s)}$

(which takes on a value of one when unit t in cluster s will occupy position j in the permutation of units within the cluster). When response is defined for all units under condition c, random variables induced by the sampling process are given by $Y_{ijck} = \sum_{s=1}^{N} U_{is} \sum_{t=1}^{M_s} U_{jt}^{(s)} Y_{stck}$.

This research will extend finite population mixed models in many directions. In nearly all cases, the target is a linear combination of random variables that arise from a finite population mixed model, such as a PSU mean, a regression coefficient, or a vector of such linear combinations. We describe extensions under simple random sampling in Section D.2, extensions in balanced two-stage cluster sampling in section D.3, extensions under two-stage sampling with unequal cluster size in Section D.4, and outline the strategy for attacking each of the problems in Section D.5. In Section D.6, we describe the time line and responsibilities.

D.2. Extensions in the Context of Simple Random Sampling (SRS)

Finite population simple random sampling offers a rich setting for extending random permutation prediction based methods. In this setting, we can view the population to be a single cluster of units. The basic model for unit t under condition c is given by $Y_{tck} = \mu_{tc} + W_{tck}$, where $E_{\xi_3}\left(Y_{tck}\right) = \mu_{tc}$ (dropping the index s). We first consider models with covariates where each unit is defined under a single condition and there is no response error, and subsequently add more than one condition, and response error. Finally, we plan to expand models that account for missing data. This extends the model $y_t = \mu_t$ (where $c = 1$ and there is no response error) investigated by Stanek, Singer and Lencina (2004), the seemingly unrelated regression results of Li (2003), and the missing data models of Lu (2004).

D.2.1. SRS Models with Covariates and No Response Error

When units are measured under a single condition ($c = 1$), inclusion of auxiliary information may lead to different models, i.e.

$$y_t = x_t\beta + \varepsilon_t$$

$$y_t = x_t\beta_t = x_t\bar{\beta} + x_t\left(\beta_t - \bar{\beta}\right) \text{ where } \bar{\beta} = \frac{1}{\mathbb{N}}\sum_{t=1}^{\mathbb{N}}\beta_t$$

$$y_t = \sum_{u=1}^{q} x_{tu}\beta_u + \varepsilon_t \text{ where } 1 < q < \mathbb{N}.$$

We will begin by developing estimators for the first two models, which naturally lead to a comparison of ratio and regression estimators in finite populations. For example, in studies of cholesterol, a common dietary measure is the percent saturated fat intake/kcal (as in the Seasons study and Watch II study), where y_t is the kcal from saturated fat, and x_t is the total kcal ingested by subject t . In the first model, assuming the average lack of fit is zero, $\sum_{t=1}^{N}\varepsilon_t = 0$, the parameter $\beta = \frac{\mu_y}{\mu_x}$ (where $\mu_x = \frac{1}{\mathbb{N}}\sum_{t=1}^{\mathbb{N}} x_t$ and $\mu_y = \frac{1}{\mathbb{N}}\sum_{t=1}^{\mathbb{N}} y_t$) has the form of a combined ratio. In the second model, β_t is the percent saturated fat (kcal/kg) for subject t , and $\left(\beta_t - \bar{\beta}\right)$ is the deviation in this value from the population average, i.e., an average of the separate ratios. Including sampling in the model, the terms $\bar{\beta} + \left(\beta_t - \bar{\beta}\right)$ for $t = 1,...,\mathbb{N}$ are latent values of randomly selected subjects. Multiplying β_t by the kcal intake for subject t results in the subject's saturated fat kcal intake, $x_t\beta_t$. The term $x_t\left(\beta_t - \bar{\beta}\right)$ is the deviation of this intake from $x_t\bar{\beta} = \frac{1}{\mathbb{N}}\sum_{t^*=1}^{\mathbb{N}} x_t\beta_{t^*}$, which represents the average kcal saturated fat intake of subjects if each subject had ingested the same amount of total kcals $\left(x_t\right)$. The ability to directly *interpret effects in these models is important, since it exposes model limitations. A limitation in the present example is analogous to Lord's paradox, where the assumption of equal kcal intake for all subjects may compare relatively high intake for small subjects with relatively low intake for heavier subjects.*

The first two models distinguish combined ratio and separate ratio parameters. In the context of a random permutation model, the term $\sum_{t=1}^{\mathbb{N}} U_{jt} x_t \left(\beta_t - \bar{\beta}\right)$ will be a random subject effect. We will investigate different settings under the assumption that x_t is known only for the sample subjects, the sample subjects values and the total of x_t are known, and settings where x_t is known for all subjects. We will use the random permutation model of Stanek, Singer, and Lencina (2004) to develop estimators of β and $\bar{\beta}$ based on a transformed model, $z_t = \frac{y_t}{x_t}$, and compare the result to estimators developed by extending the seemingly unrelated regression approach of Li (2003). These applications are particularly intriguing since the two models connect ratio parameters to random effects. We plan to develop guidelines for choice of models and estimators for common dietary measures, and relate these results to the sampling literature on ratio and regression models.

We will investigate special cases of the third model corresponding to simple linear regression i.e., $y_t = \beta_1 + x_t\beta_2 + \varepsilon_t$ and multiple regressions. As an example, let the response y_t represent the kcal intake for subject t, and x_t represent the subject's weight (in kilograms). Multiple regression models may include gender in the model, resulting in a parallel/separate slope and intercept models (which we represent by $y_t = \mathbf{X}_t'\beta + \varepsilon_t$, where $\mathbf{X}_t$ is a $u \times 1$ vector of covariates for subject t). Assuming that $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon = 0$, the parameters of interest are defined by $\beta = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$. We will develop two approaches to extend the random permutation prediction methods to the finite population regression setting. The first approach will represent each regression coefficient as a weighted sum of responses, $z_{tu} = w_{tu} y_t$, where the weights are given by $w_{tu} = \frac{x_{tu} - \mu_{tu}}{\sum_{t=1}^{\mathbb{N}}\left(x_{tu} - \mu_{tu}\right)^2}$ for $t = 1,...,\mathbb{N}; u = 1,...,q$. One approach is to stack the weighted vectors, $\mathbf{z}_u$, $u = 1,...,q$ and then consider their joint permutation using an approach similar to the seemingly unrelated regression model of Li (2003). This requires extending the results of Li and Stanek (2004) to simultaneously estimate the q parameters, β. A second approach is to directly expand the response and co-variables, $vec\left(\mathbf{y} \mid \mathbf{X}\right)$, and then use the random permutation model similar to Li (2003). We anticipate that this approach will result in estimators similar to those in Li's dissertation. Some problems are anticipated due to the zero variance of the coefficients in $\mathbf{X}$ corresponding to the intercept, which may be resolved by reducing the dimension of the random vector in the random permutation model.

Notice that when $y_t = \sum_{u=1}^{q} x_{tu}\beta_u + \varepsilon_t$ and $n < q < \mathbb{N}$, the number of parameters will exceed the sample size. This situation is similar to the situation studied by Stanek, Singer, and Lencina (2004) where $y_t = \mu_t$, and predictors of PSUs were possible by considering the parameters as random effects. We will explore framing the problem in a similar manner, and extending a similar approach the general regression model setting.

D.2.2. SRS Models with Response Error

We plan to begin by extending Stanek, Singer, and Lencina's (2004) model to include response error, i.e. $Y_{tk} = \mu_t + E_{tk}$, $k = 1,...,p_t$. If 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.

We will develop this relationship in detail by studying the limit as $M \to \infty$ of Stanek and Singer's (2004) two stage sampling results. When $p_t = 1$, the model reduces to $Y_t = \mu_t + E_t$, *which, with an external estimate of the response error variance, can be used to predict the latent value of a PSU, or the PSU mean. A natural extension of these results is to the situation where multiple measures are made on a unit. When* $p_t = p > 1$, *an estimate of the response error variance can be constructed directly from the sample data.*

We will apply these methods to problems where units' latent values are a discrete set of integers, and there is response error. For example, suppose each subject selected via simple random sampling is asked to report the number of days in the past 7 day period that s/he smoked marijuana. Suppose there is a single response for each subject. The number of days reported by the subject may differ from the actual number of days that marijuana was smoked by the subject in the past 7 days; the difference in these measures being response error. Since a subject's response would range from 0 to 7 days, and the actual number of days smoked is an integer, the response error would correspond to integer values. The response error distribution is not likely to be uniform. Considering all possible responses as a finite set, we can represent weighted response errors by expanding the number of responses. For example, for a subject who smoked marijuana on 3 of the past 7 days, we may represent the distribution of the subject's response by the *M*=9 equally likely values (1,2,2,3,3,3,4,4,5).

A selected subject may be asked to report the number of days in the past 7 days marijuana was smoked more than once in the context of a longer interview, providing multiple measures of response for the subject. Assuming the subject's response can be modeled as a simple random sample from the response distribution, then the problem corresponds to two-stage cluster sampling. Results will be extended to situations where response error sampling fractions differ between units (by design), using both with and without replacement sampling *(corresponding to multinomial or hypergeometric response error, respectively).*

We plan to develop design based methods extending of models (in Section D.2.1) where units are measured under a single condition (c=1) to include response error. Such models are given by

$$Y_t = x_t\beta + \varepsilon_t + E_t$$

$$Y_t = x_t\overline{\beta} + x_t\left(\beta_t - \overline{\beta}\right) + E_t$$

$$Y_t = \sum_{u=1}^{q} x_{tu}\beta_u + \varepsilon_t + E_t \quad \textit{where } 1 < q < \mathbb{N}$$

The first model extends Pfeffermann and Nathan's (1981) super-population regression model to include lack of fit in addition to response error in a design based framework. The second model extends Stanek, Singer, and Lencina's (2004) model to include a covariable and response error. Upon addition of sampling, the model is a single stage mixed model where the covariate can be viewed as a weighting factor for the random effects. These models can account for response error when quantifying risk factors such as kcal/kg. We will use the random permutation model of Stanek, Singer, and Lencina (2004) to develop estimators of β and $\overline{\beta}$ based on a transformed model, $Z_t = \frac{Y_t}{x_t}$, and compare the result to estimators developed by extending the seemingly unrelated regression approach of Li (2003). Although the response error variance, σ_r^2, may be initially homogeneous, in the transformed model, the response error becomes heterogeneous, $\text{var}_{\xi_3}\left(\frac{E_t}{x_t}\right) = \frac{\sigma_r^2}{x_t^2}$, *and must be accounted for in development of estimators/predictors.*

The third model includes multiple co-variables in a response error model setting. This model accounts for response error in finite population multiple regression settings. We plan on developing estimators in a similar manner as in Section D.2.1. As an example, let the response y_t represent the kcal intake for subject t, and x_t represent the subject's weight (in kilograms). Multiple regression models may include gender, resulting in a *parallel/separate slope and intercept models (which we represent by* $Y_t = \mathbf{X}_t'\beta + \varepsilon_t + E_t$*, where* $\mathbf{X}_t$ *is a* $u \times 1$ *vector of covariates for subject* t*), and* ε_t *is lack of fit. As before, assuming* $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon = 0$*, the parameters of interest are defined by* $\beta = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$*. We will develop two approaches to extend the random permutation prediction methods to the finite population regression setting. The first approach will represent each regression coefficient as a weighted sum of responses,* $Z_{tu} = w_{tu}Y_t$*, where the weights are given by* $w_{tu} = \dfrac{x_{tu} - \mu_{tu}}{\sum_{t=1}^{\mathbb{N}}\left(x_{tu} - \mu_{tu}\right)^2}$ *for* $t = 1,...,\mathbb{N}; u = 1,...,q$*. One approach is to stack the weighted vectors,* $\mathbf{Z}_u$*,* $u = 1,...,q$ *and then consider their joint permutation using an approach similar to the seemingly unrelated regression model of Li (2003). Notice that the response error variance will be heterogeneous. This requires extending the results of Li and Stanek (2004) to include response error when simultaneously estimate the* q *parameters,* β*. A second approach is to directly expand the response and co-variables,* $vec\left(\mathbf{Y} \mid \mathbf{X}\right)$*, and then use the random permutation model similar to Li (2003). This approach will extend Li's approach to include the response error for* $\mathbf{Y}$*, and account for sampling by accounting for the joint permutation of* $\mathbf{Y}$ *and* $\mathbf{X}$*.*

D.2.3. SRS Models with Units Measured under Multiple Conditions (Longitudinal Studies).

Extensions of simple random sampling models to settings where units are measured under more than one condition form a rich class of models that include longitudinal models. We plan to investigate this important class of models following an approach similar to that outlined in Sections D.2. and D.2.1. As an example, consider the WATCH II study where a pretest-posttest experimental design was used to evaluate the impact of a health systems-based referral and support program for physicians treating patients with high levels of LDL cholesterol. Nutrition classes provided an important avenue for altering LDL cholesterol. All patients (those randomized to the systems support or usual care) were eligible to attend the classes, with attendance depending upon both physician and patient factors. We develop models for the impact of the nutrition program on patients LDL measures in the context of this non-randomized pretest posttest setting. This is a simple example of a repeated measures study with $c = 1,...,2 = C$ *conditions per patients (with* $c = 1$ *for pretest, and* $c = 2$ *for posttest).*

The first model that we will investigate is given by $y_{tc} = \mu_{tc}$ *where* $\mu_{tc} = \mu + \tau_t + \varsigma_c + \varepsilon_{tc}$ *with* $\mu = \frac{1}{\mathbb{N}}\sum_{t=1}^{\mathbb{N}}\left(\sum_{c=1}^{C}\frac{\mu_{tc}}{C}\right)$ *denoting the average response (over conditions) for subjects in the population,* $\tau_t = \mu_t - \mu$ *(where* $\mu_t = \sum_{c=1}^{C}\frac{\mu_{tc}}{C}$*), denoting the deviation in average response from the mean for subject* t $\varsigma_c = \mu_c - \mu$ *(where* $\mu_c = \sum_{t=1}^{\mathbb{N}}\frac{\mu_{tc}}{\mathbb{N}}$*) denoting the deviation in average response under condition* c *from the mean, and* $\varepsilon_{tc} = \mu_{tc} - \left(\mu + \tau_t + \varsigma_c\right)$ *representing the residual for subject* t *under condition* c*. Defining* $\mathbf{y}_t = \left(y_{t1} \quad y_{t2}\right)'$*, we can directly represent sampling with the random variables* $\mathbf{Y}_j = \sum_{t=1}^{\mathbb{N}}U_{jt}\mathbf{y}_t$*,* $j = 1,...,\mathbb{N}$*, resulting in a model with two random effects (similar to a mixed model) given by* $Y_{jc} = \mu + \varsigma_c + T_j + E_{jc}$*. The two random effects in this model correspond to random subject effects, and random subject by condition interactions. Both arise directly from the probability model for sampling. The parameter* ς_c *corresponds to half the average gain. Based on the sampling model,* $\text{var}_{\xi_1}\left(\mathbf{Y}_j\right) = \sigma_\tau^2\mathbf{J}_2 + \bigoplus_{c=1}^{C}\sigma_{\tau c}^2$*, where* $\sigma_\tau^2 = \frac{1}{\mathbb{N}-1}\sum_{t=1}^{\mathbb{N}}\left(\mu_t - \mu\right)^2$ *and* $\sigma_{\tau c}^2 = \frac{1}{\mathbb{N}-1}\sum_{t=1}^{\mathbb{N}}\left(y_{tc} - \mu_c\right)^2$*. We will use the random permutation model of Stanek, Singer and Lencina (2004) to develop an estimate of the gain, and estimates of* *condition means, $\begin{pmatrix}\mu_1 & \mu_2\end{pmatrix}'$. The second set of estimators corresponding to estimating the mean response profile for subjects in the population.*

The model for response with two condition can be extended to many interesting settings. In the context of the Watch II study, some subjects attend nutrition clinics ($x_t = 1$), while others do not ($x_t = 0$). Accounting for such non-random clinic attendance is an example of inclusion of a covariable in the model. This covariable may be related to average subject's response at pretest (since the average response of subjects who will attend the nutrition clinics may have higher LDL levels), and may also be related to response at posttest (since there may be a smaller gain in response for subjects attending the nutrition clinics). We will extend the model to account for such a covariable.

A second extension corresponds to more than two conditions. For example, in many studies, subject's response is recorded at multiple time points (such as baseline, 6 months, 12 months, and 18 months. With such multiple measures, in addition to response profiles, we may consider reduced models that reflect simple linear growth over time (where $x_{c1} = 1$ and $x_{c2} = (c-1)$), i.e., $\mu_c = \sum_{u=1}^{q=2} x_{cu}\beta_u$ (and $q = 2$). These models may be expanded to include population parameters and individual subject parameters (where $q = 2\mathbb{N}$) as well as lack of fit, such that $y_{tc} = \beta_1 + x_{c2}\beta_2 + (\beta_{t1} + x_{c2}\beta_{t2}) + \varepsilon_{tc}$, where $\sum_{t=1}^{\mathbb{N}} \beta_{tu^*} = 0$ for $u^* = 1,2$. This is a simple example of a growth model. We will develop estimators of population parameters, and predictors of random effects based on random variables resulting from a random permutation model based on simple random finite population sampling. We anticipate that in this situation (due to x_c not depending on t), it will be possible to separate the design matrix in the prediction based approach from subject effects, and thus avoid the problem of a random design matrix that was a limitation of Porter's (1973) development.

An important extension of this problem is to settings where subjects are observed under different conditions (such as when the time points for response differ between subjects). This problem is more complicated than the previous settings since both cross-sectional and longitudinal effects are possible. For example, suppose that conditions correspond to a subject's age, with ages given by $c = 1,...,C = 4$, with the ages given respectively as $12, 13, 14,$ and 15. Furthermore, suppose that although population parameters are defined as if response for each subject is potentially observable at all ages, some subjects may only be observed at age 12 and 13, while other subjects may only be observed at age 14 and 15. This corresponds to a setting, similar to the Urban City Schools study, where observations are made on 6th and 8th grade students in a given school year. Cross-sectional effects may correspond to an increase in response (such as bullying in school) from grade 6 (12/13 year old students) to grade 8 (14/15 year old students), where as longitudinal effects may correspond to changes between age 12 and 13 (and age 14 and 15) in response. This is the simplest setting where tracking of cross-sectional and longitudinal effects is critical. We will develop appropriate parameterizations of the models that have practical interpretations, and use these models to expand the development of estimators based on the random permutation model.

Additional extensions of these models will be considered when there is response error. Although we do not detail such extensions here, they will build upon results from Section D.2.1. Models with response error and lack of fit will enable expressions for estimators/predictors to distinguish the relative contribution of these variance components, similar to the development by Stanek and Singer (2004).

D.2.4. Extensions to Settings where there and Multiple Responses, and Possibly Missing Response

We plan to extend each of the models described in sections D.2, D.2.1, D.2.2, and D.2.3 to settings where there are multiple measures on a subject under a condition $(p_{st} > 1)$. In such settings, it is possible to pool squared deviations of response to estimate response error. We will consider settings where each subject has *the same number of measures* $\left(p_{st} = p > 1\right)$*, and the setting where these numbers differ by design. In addition, we will investigate the setting where by design, an equal number of measures are to be made, but some measures are missing.*

For example, 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.

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_{tk} that takes on a value of 0 (if missing) or 1 for each potential response. Using this idea, data for response *k* for subject *t* can be viewed as a bivariate pair, $Y_{tk}X_{tk}$, X_{tk} , where X_{tk} follows a Bernoulli (P). The sample data is generated by a two-stage process: in the first stage, subjects are sampled, while in the second stage, the bivariate pairs are selected and the corresponding response is measured (or not) according to the realization of X_{tk}. Such a process underlies data that are missing at random. A prediction-based approach will be used to estimate linear functions of the corresponding random variables. *This approach is will build on the work of Lu (2004).*

Other frameworks that represent the missing data process may be conceived. For example, the probability that response is missing may be related to an interviewer, and hence to the assignment of selected subjects to an interviewer. In such a setting, we will represent the missing indicator random variable as $Y_{jk} = \sum_{t=1}^{\mathbb{N}} X_{jk} U_{jt} Y_{tk}$ *, similar to Lu (2004).* We will investigate such frameworks and develop basic results in the simplest settings. Extensions will be made to settings where the Bernoulli parameter depends in some fashion on *either the subject or sample position*.

D.3. Extensions in the Balanced Two-stage Sampling

In populations with equal size clusters, we will extend results to settings where response is measured more than once for a sampled SSU. We will assume that the number of responses for each SSU is specified by the study design, and not by a missing data mechanism. This is the basic step that will lead to methods for hierarchical models with more than two levels. We will begin by assuming that the same number of response measurements is obtained on each selected SSU.

Development of predictors of PSU means in this setting will be made in a similar manner as the development outlined in section C.2.1. and C.2.2.. Of particular interest is the application to two stage cluster sampling with response error where, for example, clusters correspond to groups of subjects (such as students in a classroom) and response corresponds to a behavior (such as smoking marijuana) on a given day, with the subject's parameter defined as the average response over a specified time period (such as a week or month). We will consider other hierarchical model settings where target parameters correspond to the mean response for a unit at a specified level of the hierarchy.

D.3.1. Two-stage Sampling with Cluster Level Covariates and No Response Error

A simple finite population setting that includes covariates occurs when a simple random sample is selected and the population is composed of domains (formed from age groups or gender). This problem has been investigated by (Li 2003) in a recently completed Ph.D thesis (section C.5.). We plan to extend the results to the two-stage cluster sampling problem. Considering time to be a SSU dependent covariate will link this approach to the development outlined in Section D.2.3. First, we will consider cluster level covariates.

As an example, consider a setting where clusters are physician's practices (clinics) and units are the patients of the physician. The cluster-level covariates may include the physicians' specialties, gender, race, years of practice, affiliation to a medical school, languages spoken, etc. and be categorical (gender, race, specialty) or continuous (age, years of practice). We propose to extend the method developed by (Li 2003) to the two-stage random permutation model presented by (Stanek and Singer 2004). Briefly, we first represent the two stage permutations for the response variable and covariate(s) jointly using a seemingly unrelated regression set-up to link the permutations of the vectors of response variable and covariates. This will define a two-stage random permutation seemingly unrelated regression model.

We sketch the approach for seemingly unrelated regression models in a population formed by equal size clusters with no response error. We represent the $M \times 1$ response vector for cluster s as $\mathbf{y}_s$, and the corresponding vector of auxiliary values as $\mathbf{x}_s$. The response vector for the entire population is denoted as $\mathbf{y} = \left(\left(\mathbf{y}_s \right) \right)$, and the vector of auxiliary variable as $\mathbf{x} = \left(\left(\mathbf{x}_s \right) \right)$. Defining $\mathbf{U}$ as an $N \times N$ permutation matrix as in C.3, $\mathbf{Y} = \left(\mathbf{U} \otimes \mathbf{I}_M \right) \mathbf{y}$ represents a permutation of cluster response vectors, while $\mathbf{X} = \left(\mathbf{U} \otimes \mathbf{I}_M \right) \mathbf{x}$ represents the corresponding permutation of cluster auxiliary vectors. We form a seemingly unrelated regression model by concatenating these two vectors such that $\begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} = \mathbf{X}^{\circ} \boldsymbol{\beta} + \mathbf{E}$, where $\mathbf{X}^{\circ}$ is the design matrix for the model (such as $\mathbf{I}_2 \otimes \mathbf{1}_{NM}$). Additional auxiliary variables can be added to the model, and alternative design matrices can be explored. Following the method of (Li 2003), we will then derive linear unbiased minimum variance predictors of response population means and totals by applying Royall's general prediction approach (Royall 1973; Valliant, Dorfman et al. 2000).

When only cluster-level covariates are included in the model, there will be no variation for the covariate within each cluster. For this setting, only the first stage permutation is needed for the covariate and the derivation is anticipated to be significantly simpler *(as in Section D.2.1).*

D.3.2. Two-stage Sampling with Multiple Response

We plan on developing predictors where, by design, response is measured a different number of times for different sampled SSUs. For example, in studies of arsenic ingestion from food, a duplicate food sample is collected on a day (SSU) for a subject (PSU), with a sub-sample of the food analyzed for arsenic content. Due to high analysis costs, a second sub-sample will be analyzed on only a fraction of selected SSUs. The sub-samples form a planned third level in the hierarchy. In this setting, we will investigate differences between predictors of the latent value of a PSU and the average value of the PSU (as discussed in (Stanek and Singer 2004)).

D.3.3. Two-stage Sampling with Multiple Conditions (Longitudinal Studies, Random Coefficient Models)

We begin by considering the simplest context, where a population is composed of clusters *of units (such as patients in clinics, or students in schools),* and the unit is measured under two conditions *(baseline a 1 year follow-up)*. A cluster mean is defined as the average of the expected response at two fixed times. For example, response may be measures of patient's fasting total serum cholesterol at baseline, and 1 year follow-up. Estimates of the change in cholesterol for clinics, and estimates of the average change in the population are of interest. *This is a simple example of a repeated measures study in a hierarchical population with $c = 1,...,2 = C$ conditions per patients (with $c = 1$ for pretest, and $c = 2$ for posttest).* Examples such as this are common in health sciences.

We will build models in this setting following a similar plan as in Section D.2.3. *The first model that we will investigate is given by* $y_{stc} = \mu_{stc}$ *where* $\mu_{stc} = \mu_s + \tau_{st} + \varsigma_{sc} + \varepsilon_{stc}$ *with* $\mu_s = \frac{1}{M}\sum_{t=1}^{M}\left(\sum_{c=1}^{C}\frac{\mu_{stc}}{C}\right)$ *defined as the average response (over conditions) for subjects cluster* s, $\tau_{st} = \mu_{st} - \mu_s$ *defined as the deviation in average response from the cluster mean for subject* t *(where* $\mu_{st} = \sum_{c=1}^{C}\frac{\mu_{stc}}{C}$*),* $\varsigma_{sc} = \mu_{sc} - \mu_s$ *defined as the deviation in average response of cluster* s *under condition* c *(where* $\mu_{sc} = \sum_{t=1}^{M}\frac{\mu_{stc}}{M}$*) from the cluster mean, and* $\varepsilon_{stc} = \mu_{stc} - \left(\mu_s + \tau_{st} + \varsigma_{sc}\right)$ *representing the residual for subject* t *in cluster* s *under condition* c*. Defining* $\mathbf{y}_{st} = \left(y_{st1} \quad y_{st2}\right)'$*, we can directly introduce two-stage cluster sampling with the random variables* $\mathbf{Y}_{ij} = \sum_{s=1}^{N} U_{is} \sum_{t=1}^{M} U_{jt}^{(s)} \mathbf{y}_{st}$, $i = 1,\ldots,N;\, j = 1,\ldots,M$ *. Introduction of the indicator random variables in this model will produce random effects, such as the random PSU effect (i.e.,* A_i*) given by* $\sum_{s=1}^{N} U_{is}\mu_s = \mu + \sum_{s=1}^{N} U_{is}\left(\mu_s - \mu\right) = \mu + A_i$*, or a random effect representing the gain (i.e.* $\varsigma_{sc} = \mu_{sc} - \mu_s$*) for PSU* i *. We will use the two-stage random permutation model of Stanek and Singer (2004) to develop a estimators of fixed effects and predictors of random effects in the model.*

This approach will be generalized to three time points, and to p time points, with response recorded for each SSU at each time point. This gives rise to growth curve models for SSUs over time, and will be used to develop predictors in finite population random coefficient regression models. These models will account for lack of fit for units, and distinguish lack of fit from the variance of units in a cluster.

D.3.4. Two-stage Sampling with Multiple Conditions and Unit Covariables

We will develop predictors for two-stage sampling models with repeated measures of units over time where an additional covariable is present for a unit. As an example, gender (the covariable) may be recorded for each patient (unit) in a cluster (clinic), with response for the patient reported under different conditions (times). This extends models for domains in a simple random sampling settings (as discussed by Li and Stanek (2004)) to two-stage sampling settings where the domains are defined for the units, and further extends these models to situations where units are measured under different conditions (time).

This problem is complex, but non-stochastic models for response can be clearly defined for units in the population. By adding sampling indicator random variables for these models similar to Stanek and Singer (2004), parameters corresponding to fixed effects and random effects can be defined, and estimators/predictors can be developed.

This problem can be further extended in a similar manner to the models described in Section D.2.3. where the conditions (time points) differ between subjects. There are many variations of such models, where the different models are characterized by different main effects and interactions, and there are possibly longitudinal and cross-sectional trends. We plan to develop such models in the context of research studies (such as the Seasons Study or Urban City Schools study) so as to focus the model choice of main effects and interactions, and the solutions.

D.4. Extensions to Two-stage Sampling with Unequal Size Clusters

We plan to study in more depth methods for prediction in populations formed from unequal size clusters by focusing on the synthesis of principles underlying expansion and projection of random variables, as described in C.3. This study will draw upon results from the related literature concerning ancillary variables, and orthogonal projections of nuisance parameters and apply them to the finite population mixed model setting.

Predictors of PSU means have been developed by (Stanek and Singer 2004) in populations formed by clusters of unequal size, as discussed in C.4. Development of these predictors requires specifying an expanded set of random variables to represent a permutation of the population, as indicated in C.3. This expansion is required because of a limitation of the standard representation of random variables, where subscripts correspond to positions (ij) rather than labels (st) of the units in the population. The expanded model extends the typical permutation model to a broader set of random variables, but falls short of a model for the very general set of random variables envisioned by (Godambe 1955) which (in the context of single stage sampling) span an $(N-1)^n$ dimensional space. The random variables in a typical permutation model span an $N-1$ dimensional space. The random variables in the expanded model span an $(N-1)^2$ dimensional space. Models with fewer random variables than Godambe's model, but more random variables than the expanded models may be postulated that may lead to new insights.

We plan to explore this area to better understand and define the strategies for developing finite population mixed models. The investigation will pursue two avenues of research. These two directions are aimed at the two unusual features of the development: expansion of the representation of random variables under the two-stage permutation model, and projection of the expanded random variables to a space where unique estimation is possible. More details are given in technical reports on the cluster Web site.

D.5. Additional investigations.

As time permits, we plan to explore additional areas. We anticipate that there will not be adequate time to develop these areas in depth, but if other work proceeds more rapidly than expected, we plan to initiate research in these areas. One area to be developed includes settings where the cluster mean is a multiplicative function of the population mean. The first step in this development will be a review of the literature for multiplicative models applied to the finite population context. This review will include the (Grizzle, Starmer and Koch 1969) approach implemented in SAS Catmod, as well as the estimating function approach outlined in (Thompson 1997). 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 and Hauck 1991) and discussed by (Holt 1989). We will identify possible design based approaches 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 be exploratory and attempt to adapt ideas from this literature to the finite population mixed model. Other areas that will be considered for development will include response error, longitudinal response, and covariables in two-stage cluster sampling models with unequal cluster size. We will also explore expanding the missing data representation of Lu (2004) to two-stage sampling settings, and unequal probability sampling.

D.6. Strategy for Attacking the Problems

We plan to follow a similar strategy to develop and evaluate estimators/predictors in each area. First, we will define units in the population, response variables, covariables, and the structure of these units. These definitions will be made in the context of one of the research studies that help motivate these methods. Population parameters and unit parameters will be defined. We will then use random permutation models to specify the permuted population, partition the population into the sample and remaining units, and develop optimal estimators/predictors along with their EMSE. The developed predictors will be compared with predictors from the sampling and mixed model literature. Empirical predictors will be constructed by replacing variance components by their estimates, and we will develop approximations for the EMSE of the empirical predictors. We anticipate that these predictors can be constructed using variance component estimates. We will develop simple SAS and STATA code (as macros) that can be used to practically implement the results.

Finally, we will simulate finite populations with different underlying distributions to evaluate the predictors and empirical predictors, both confirming the theoretical results, and evaluating the results of the empirical *predictors. The simulation studies will consider a wide range of settings, but also be focused on settings similar to those encountered in applications.* We will collaborate with epidemiologists and cardiologists in refining variables and risk factors to maintain the relevance of the simulations. Dr. Elizabeth Bertone (a nutritional epidemiologist), Dr. Lisa Chasan-Taber (an exercise epidemiologist), and Dr. Ira Ockene (a cardiologist) will provide advice on parameter definition and relevant time periods based on knowledge of specific disease outcomes, plausible interventions, and practical considerations. *All work will be developed into technical reports posted on the WEB and refined in manuscripts in peer reviewed journals.*

D.6.1. Development of Estimators and Predictors

We outline the general strategy to be used for developing estimators and predictors. The basic features of the approach are given in the papers by Stanek, Singer and Lencina (2004) and Stanek and Singer (2004), and reviewed in Section C.2. First, we will clearly define the population including units and levels of units, conditions, and response variables. Next, we will define finite population parameters for the models. Apart from response error, all terms in the definitions are non-stochastic. We will formulate these models using standard matrix algebra for the units in the population. Sampling random variables will be introduced using the indicator random variable definitions as in Stanek and Singer (2004). An important property of the models is a definition that results in non-stochastic design matrices for the fixed and random effects. This feature may be obtained by transforming response (such as by dividing by the value of a covariate), by multiplying by some non-stochastic orthogonal matrix (similar in spirit to transformations applied by Fuller and Battese (1973), or adapted from Pfeffermann and Nathan (1981)), or by simple subtraction of a known value (as in Li and Stanek (2004)). When the design matrices are non-stochastic, development of the predictor is straightforward (avoiding the problems noted by Pfeffermann (1984) in Porter's (1973) development).

Parameters and random variables of interest will be defined that are linear combinations of the permuted random variables. We will focus development of predictors of parameters defined in terms of latent values. The expected value and variance of the permuted random variables will be developed, paying special attention to defining simple notational conventions. Predictors will be developed by partitioning the random variables into a sample and remainder, defining unbiased criteria, and minimizing the resulting expression for the EMSE. We plan to minimize the trace of the MSE for parameter vectors. Our focus is on parameters that are a linear combination of the random variables. Since parameters and their estimators/predictors are linear functions of the stochastic population, it will be possible to develop explicit expressions for the EMSE.

D.6.2. Comparison of Predictors

Using the random permutation model variance, we will compare the EMSE with the EMSE of other predictors from the sampling and mixed model literature. Although the predictors developed in this research will be optimal, in some settings other predictors commonly in use will have EMSE that come close to the minimum. We will use these comparisons to identify where similar results may be expected, and where the results will be different, patterning this development after Stanek and Singer (2004). In the simplest setting, the comparisons will depend on values of *N, n,* M_s , m_s , σ^2 ; σ_e^2 ; and σ_r^2 . As a result of these efforts, we will gain an understanding of settings where predictors differ substantially, and where large reductions in the EMSE occur. This knowledge, plus the applications, will be used in designing the range of characteristics for the simulation studies. The evaluation of the proposed predictors involves replacing the shrinkage factors by estimates leading to what we refer to as empirical predictors; developing analytic expressions for their bias, variance and MSE; developing interval estimates; and conducting simulation studies to evaluate their performance.

D.6.3. Empirical Predictors

*Empirical predictors will be constructed by replacing variance components by their estimates. Since a*ll of the predictors developed for the finite population mixed model to date include shrinkage factors that are a non-linear function, we plan to estimate these shrinkage factors using moment estimates similar to those arising in predicting random effects in the usual mixed model (see for example Sections C.2.2 and C.4). Some v*ariance* *components, such as σ_s^2, can be directly estimated by sample moments. Other components may no be so easily estimated, and we explore estimators constructed from sample moments, or direct use of restricted maximum likelihood estimators from normal based mixed models. We also will explore a functional approach for empirical estimation, building on Binder's (1983) work on variances of asymptotically normal estimators from complex surveys.*

We plan to develop methods of estimating the EMSE of the empirical predictors. We will pursue two approaches for this development. The first approach will use a Taylor series expansion of the predictor to express it in terms of a linear function of first and second sample moments, followed by direct evaluation of the approximated EMSE. This will involve fourth order sample moments. A second approach will use an approximation given by (Prasad and Rao 1990) and applied to a random effects model (Prasad and Rao 1999). For predictors developed under a mixed model or super-population model (Stukel, Hidiroglou et al. 1996; Duchesne 2000; Valliant, Dorfman et al. 2000; Valliant 2002), we will also compare the EMSE derived under the finite population mixed model with other expressions of the EMSE available in the literature (Kleffe and Rao 1992). We will begin with populations formed by equal size clusters, add response error, and then progress to populations formed by unequal size clusters.

D.6.4. Development of Software for Empirical Predictor

We will develop additions to SAS (and STATA) software that will produce the empirical predictors developed in the research. The computer code will enable other researchers to account for finite population characteristics when fitting mixed models in hierarchical settings. We anticipate developing macros in SAS that will take as input data on population characteristics (such as cluster sizes), and results from other SAS procedures (such as means, totals, or Proc Mixed results), and produce output similar to that produced by PROC Mixed for random effects, or regression parameters. We plan to use the Cluster WEB site (http://www.umass.edu/cluster/) to distribute the software macros to other researches. This website was established in 2000 to support and serve as an interchange for technical material, manuscripts and software related to analysis of finite population clustered data. The website will illustrate the use of the macros with examples taken from research studies, provide links to manuscripts describing the methods, and provide links to a simulation study site that enables practical evaluation of the properties of the predictors. We also plan to use the website to interchange material and results as part of this research.

D.6.5. Simulation Studies

A primary goal of the simulation studies is the evaluation of the performance of the finite population mixed model methods relative to alternative predictors based on standard mixed models, and super-population models (Scott and Smith 1969; Bolfarine and Zacks 1992). In the past year, we have developed a simulation for two stage cluster sampling with response error, and are currently in the process of using it to study the properties of empirical predictors of PSU means (see Section C.6). The simulation compares the corresponding bias, the theoretical EMSE, and the estimate of the empirical predictor EMSE between different predictors, allowing for different distributions of cluster means, and distributions of units with the clusters.

A secondary goal of the simulation is the evaluation of interval estimators of target random variables based on normal approximations for 95% confidence intervals using the square root of the estimated empirical EMSE. The actual coverage, width, percentages of left- and right-misses of interval estimators will be evaluated. Currently, the simulation does not evaluate coverage of interval estimates, even using theoretical EMSE of the predictors. A module will be added to the simulation to evaluate coverage.

D.6. Time line and responsibilities

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. Doctoral students at UMASS and USP will have an active role in the process. The activities will be coordinated by Dr. Stanek at UMASS, and Dr. Singer in Brazil. Previous collaborations have been facilitated by development and maintenance of an extensive research WEB site, frequent communication via Email, and bulletin boards and face to face working meetings (of one or two weeks duration each year). This research will use a similar strategy to maintain and enhance collaboration. Web and Email communication will be augmented by two working meetings each year (Jan. 2 weeks at USP), and July (4 weeks at UMASS). Drs. Singer and a doctoral student from USP will attend the meetings at UMASS, and Drs. Stanek, Li, Reed and a doctoral student will attend the meetings in Brazil each year. The UMASS meetings will also be joined by Drs. Bertone, Chasan-Taber, and Ockene, who will provide input and context for applications. Dr. Bolfarine will attend meetings at USP each year. The research WEB site will catalogue developing manuscripts and reports, identify current directions and ideas, and document interactions. The WEB site will be used to promulgate research results that will be presented at national meetings and peer reviewed journals. We also plan to develop a portion of the site for applications, and develop material to facilitate access for other researchers. The WEB site will be promoted at presentations National and International meetings.

Individual investigators will take the lead on particular efforts, with a primary and *secondary* collaborator. We will maintain an emphasis on building results from simple to more complex settings. Working project meetings will be used to share results, and enhance additional investigator collaborations. During the first year of the project, we plan to conduct the analytic work on EMSE (Reed, *Stanek*) necessary to design and implement the simulation studies that will evaluate the performance of the finite population mixed model predictors (Reed, *Li*, Stanek). Along with this development, we plan to extend the models to simple random sampling designs with measurement error, and investigate ways of incorporating missing data mechanisms into the predictors (Singer, *Stanek, and* Li*, Stanek*). We also will begin a more in depth review of the literature aimed at resolving methodological issues related to populations formed by unequal size clusters (Stanek, *Singer*, Bolfarine). Manuscripts and presentations will be prepared and submitted to peer reviewed journals and national and international meetings. In the second year of the project, we plan to continue simulation studies, applying them to new predictors as they emerge (Reed, *Li*). We also plan to develop applications on domain estimation (Li , *Stanek*) to two stage cluster settings and cluster level covariates (Stanek, Li). Work will continue on resolving methodological issues, and build upon the earlier years results (Stanek, *Singer*,). Doctoral students at UMASS and USP will participate in this process. Work on longitudinal applications will be initiated (Singer, *Stanek*, Li, Reed). The results of this work will be developed into manuscripts and presentations and submitted to peer reviewed journals and national and international meetings. In the final year of the project, we will continue evaluating new predictors, and developing applications to epidemiological settings. One focus will be an increased emphasis on manuscripts that illustrate application of the methods. We will build upon the result of earlier years, and extend such results where possible. 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.

E. Human Subjects Research

This research will evaluate and develop statistical methods for analyzing data from finite clustered populations. No data will be collected on any subjects as part of the study. Only existing data sets will be used in the study, and such data sets will have information recorded such that subjects cannot be identified, directly, or through identifiers linked to the subjects. The use of such existing data sets will be for guidance in selection of parameters for simulation studies, or for illustration of the methodology. We plan to use data that has resulted from *five* studies:

1. The Seasons Study (NHLBI:R01-HL52745. Ockene, I PI)

This study enrolled a volunteer sample of 641 subjects from the Fallon Health Maintenance organization over a 4 year time period. The study was conducted by the UMASS Medical center. Subjects participating in the study completed informed consents, and the study was approved by IRB-human subjects committees. Each study participant was followed quarterly for 5 quarters, with dietary, physical activity, and cholesterol data recorded. To be eligible, subjects had to be over 18 years of age, and not on any dietary restrictions or cholesterol lowering medications. An effort was made to enhance participation by black subjects. Data are stored in SAS data sets will all subject identifiers removed.

2. The Pathways Study (R01-DS11421. Schensul, J. PI))

This study enrolled a volunteer group of 401 youth between the ages of 16 and 24 in a longitudinal study of progression in substance use. The study was conducted by the Institute of Community Research, Hartford. Subjects participating in the study completed informed consents, and the study was approved by IRB-human subjects committees. Each subject was interviewed two times separated by a one year period as part of the study. Measures of self-reported substance use was recorded over different recall time periods (1 day, 30 days, lifetime) in the study. Enrolled subjects were primarily black and Hispanic. Data are stored in SAS data sets will all subject identifiers removed.

3. The Physical Activity in Pregnancy Study (R03-HD-393441, Chasan-Taber, L. PI)

This study performed 24-hour physical activity recalls among 262 racially diverse and predominantly low income prenatal care women aged 16 to 40 at a large tertiary care facility in Western Massachusetts. Subjects participating in the study completed informed consents, and the study was approved by IRB-human subjects committees. The goal of the study was development of a physical activity instrument for pregnant women. Subjects were asked to account for the minutes of each one-hour period from 12:00 midnight through the following 24 hours. Activities were classified by intensity in MET equivalents. Data are stored in SAS data sets will all subject identifiers removed.

No subject identifiers will be present in any data used. In addition, data are kept confidential on a secure password protected local area network system. As a result of these considerations, this research qualifies for Exemption 4 (page 25 of NIH Section 1, Preparing Your Application 1/13/2003).

4. Building Preventative Group Norms in Urban Middle Schools (R01-DA12015, Schensul, Jean J. PI)

This study recruited a school district in a large urban inner city, and participation by 12 middle schools in the district. Middle schools were pair matched by student characteristics (size, SOES, and racial/ethic distribution) and one school in each pair was randomly assigned to an intervention consisting of a health curriculum that developed individual and group norms resistant to drug use. A questionnaire was administered to nearly all (~85%) of students in selected grades annually in the school system. One cohort (Cohort A) of students had no intervention (6th grade students in 1999-2000). Teachers in the intervention group were trained in the intervention, and delivered the intervention (in classrooms) to 6th grade students (Cohort B) in the spring semester of 2000-2001, and once again when the students were in 7th grade in 2001-2002. A third cohort

(Cohort C), students who were in 6th grade in 2001-2002) received the intervention in 2001-2002 and in 2002-2003. Follow-up continued on students through 8th grade. A variety of measures of student behavior, perceptions, and skills were assessed as part of the questionnaire.

5. WATCH II Study (R18-HL44492-05A1, Ockene, I. PI)

The Worcester Area Trial to Reduce Cholesterol (WATCH II) study implements and evaluates the effects of a systems-based nutritional intervention program for the patients of primary care internists on dietary intake of saturated fatty acids (SFA) and on serum low density lipoprotein cholesterol levels (LDL-C). In this two-condition randomized clinical trial, the Intervention condition is the systems-based intervention added to the previously developed and found-to-be-efficacious physician-delivered nutrition intervention training and office support program (the control condition). Baseline measures and 1 year follow-up measures are collected on patients in the trial.

Inclusion of Women

The gender distribution for the studies is as follows:

Study	Number Female (Percent Female)
Seasons Study	310 (48%)
Pathway Study	117 (29%)
Physical Activity in Pregnancy Study	262 (100%)
Urban Middle School Study	3273 (51%)
WATCH II Study	398 (55%)

Only women were eligible for the Physical Activity Study.

Ethnic Distribution

Study	Number Hispanic (Percent Hispanic)
Seasons Study	24 (4%)
Pathway Study	214 (53%)
Physical Activity in Pregnancy Study	64 (28%)
Urban Middle School Study	1758 (27%)
Watch II Study	18 (2%)

The Seasons study had few Hispanic participants due to the low number of Hispanics in the Worcester area community.

Racial Distribution

Study	Am. Indian	Asian	Black	Hawaiian	White
Seasons	0	11 (2%)	85 (13%)	0	545 (85%)
Pathway	8 (2%)	1 (0.2%)	381 (95%)	0	11 (3%)
Pregnancy	2 (1%)	5 (2%)	97 (42%)	0	125 (55%)
Urban Sch.	1 (<1%)	93 (1%)	3659 (57%)	0	2669 (41%)
Watch II	8 (1%)	10 (1%)	34 (4%)	0	635 (89%)

The Pathways study did not distinguish ethnicity clearly from race. We report people who classified themselves as Hispanic as Black.

Inclusion of Children

Study	Number Under 21 (Percent Under 21)
Seasons Study	0 (0%)
Pathway Study	286 (71%)
Physical Activity in Pregnancy Study	43 (19%)
Urban Middle School Study	6472 (100%)
Watch II Study	0 (0%)

The demanding nature of follow-up in the Seasons study resulted in no subjects under age 21. The Watch II study also had no subjects under age 21.

G. Literature Cited

(including several references not cited)

Basu, D. (1958). "On sampling with and without replacement." Sankhya:The Indian Journal of Statistics: Special issue on Sample Surverys 20: 287-294.

Basu, D. and J. K. Ghosh (1967). "Sufficient statistics in sampling from a finite population." Bulletin of the International Statistical Institute 42(2): 85-89.

Bellhouse, D. R. (1987). "Model-based estimation in finite population sampling." American Statistician 41: 260-262.

Bellhouse, D. R. and J. N. K. Rao (2002). "Analysis of domain means in complex surveys." Journal of Statistical Planning and Inference 102: 47-58.

Bellhouse, D. R., M. E. Thompson and V. P. Godambe (1977). "Two-stage sampling with exchangeable prior distributions." Biometrika 64(1): 97-103.

Binder, D.A. (1983). "On the variances of asymptotically normal estimators from complex surveys." International Statistics Review 51:279-292.

Bolfarine, H., L. B. Gasco and P. L. Iglesias (2003). "Inference under representable priors for Pearson type II models in finite populations." Journal of Statistical Planning and Inference.

Bolfarine, H. and S. Zacks (1992). Prediction Theory for Finite Populations. New York, Springer-Verlag.

Brewer, K. R. W. (1979). "A class of robust samling designs for large-scale surveys." Journal of the American Statistical Association 74: 911-915.

Brewer, K. R. W. (1995). Combining design-based and model-based inference. Business survey methods. Cox, Binder, Chinnappaet al. New York, John Wiley and Sons: 589-606.

Brewer, K. R. W. (1999). "Cosmetic calibration with unequal probability sampling." Survey methodology 25: 205-212.

Brewer, K. R. W. (1999). "Design-based or prediction-based inference? Stratified random vs stratified balanced sampling." International Statistical Review 67: 35-47.

Brewer, K. R. W., M. Hanif and S. M. Tam (1988). "How nearly can model-based prediction and design-based estimation be reconciled?" Journal of the American Statistical Association 83: 128-132.

Brewer, K. R. W.(2002). Weighing Basu's Elephants. New York, Oxford University Press.

Brown, H. and R. Prescott (1999). Applied mixed models in medicine. New York, John Wiley and Sons.

Cassel, C. M., C. E. Sarndal and J. H. Wretman (1976). "Some results on generalized difference estimation and generalized regression estimation for finite populations." Biometrika 73: 615-620.

Cassel, C. M., C. E. Sarndal and J. H. Wretman (1977). Foundations of inference in survey sampling. New York, John Wiley and Sons.

Cochran, W. (1977). Survey Sampling, John Wiley.

Cox, D. R. and N. Reid (1987). "Parameter orthogonality and approximate conditional inference." Journal of the Royal Statistical Society B 49(1): 1-39.

Crowder, M. J. and D. J. Hand (1990). Analysis of repeated measures. New York, Chapman and Hall.

Datta, G. S. and J. K. Ghosh (1995). "On priors providing frequentist validity for Bayesian inference." Biometrika 82(1): 37-45.

Datta, G. S. and M. Ghosh (1991). "Bayesian prediction in linear models: applications to small area estimation." Annals of Statistics 19: 1748-1770.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm." Journal of the Royal Statistical Society B 39: 1-38.

Deville, J.-C. and C.-E. Sarndal (1992). "Calibration estimators in survey sampling." Journal of the American Statistical Association 87(418): 376-382.

Diggle, P. L., P. Heagerty, K. Y. Liang and S. Zeger (2002). Analysis of Longitudinal Data, Oxford University Press.

Duchesne, P. (2000). "A note on jackknife varaince estimation for the general regression estimator." Journal of Official Statistics 16: 133-138.

Efron, B. (1979). "Bootstrap methods: another look at the jackknife." Annals of Statistics 7: 1-26.

Efron, B. and R. J. Tibshirani (1993). An introduction to the bootstrap. New York, Chapman and Hall.

Eisenhart, C. (1947). "The assumptions underlying the analysis of variance." Biometrics 3: 1-21.

Ericson, W. A. (1969). "Subjective Bayesian models in sampling finite population I." Journal of the Royal Statistical Society B 31: 195-234.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations: Stratification. New developments in survey sampling. N. L. J. a. H.Smith. New York, Wiley-Interscience: 326-357.

Fuller, W. A. and G. E. Battese (1973). "Transfomrations for estimation of linear models with nested-error structure." Journal of the American Statistical Association 68(343): 626-632.

Gao, D. D. and R. B. Huang (2000). "Some finite sample properties of Zellner's estimator in the context of m seemingly unrelated regression equations." Journal of Statistical Planning and Inference 88: 267-283.

Ghosh, M. and G. Meeden (1986). "Empirical Bayes estimation in finite population sampling." Journal of the American Statistical Association 81(396): 1058-1062.

Ghosh, M. and G. Meeden (1997). Bayesian methods for finite population sampling. New York, Chapman Hall.

Godambe, V. P. (1955). "A unified theory of sampling from finite populations." Journal of the Royal Statistical Society B 17: 269-278.

Godambe, V. P. (1970). "Foundations of survey sampling." The American Statistician 24: 33-38.

Godambe, V. P. (1991). "Orthogonality of estimating functions and nuisance parameters." Biometrika 78(1): 143-151.

Godambe, V. P. and M. E. Thompson (1989). "An extension of quasi-likelihood estimation." Journal of Statistical Planning and Inference 22: 137-152.

Goldberger, A. S. (1962). "Best linear unbiased prediction in the generalized linear regression model." American Statistical Association Journal 57: 369-375.

Goldstein, H. (2003). Multilevel statistical modeling, 3rd Edition. London, Kendall's Library of Statistics 3. Arnold.

Graybill, F. A. (1983). Matrices with applications in statistics. Belmont, California, Wadsworth International.

Grizzle, J. E., C. F. Starmer and G. G. Koch (1969). "Analysis of categorical data by linear models." Biometrics 25: 489-504.

Hansen, H. M., G. W. Madow and J. B. Tepping (1983). "An Evaluation of Model-dependent and Probability-Sampling Inferences in Sample surveys." Journal of the American Statistical Association 78(384): 776-793.

Hansen, M. H., T. Dalenius and J. B. Tepping (1985). The development of sample surveys of fintie popualtions. A Celebration of Statistics. A. C. Atkinson and S. E. Fienberg. New York, Springer-Verlag: 327-354.

Hanurav, T. V. (1966). "Some aspects of unified samling theory." Sankhya Series A 28: 175-204.

Hanurav, T. V. (1968). "Hyper-admissibility and optimum estimators for sampling finite populations." Annals of Mathematical Statistics 39: 621-642.

Hartley, H. O. and J. N. K. Rao (1968). "A new estimation theory of sample surveys." Biometrika 55: 547-557.

Hartley, H. O. and J. N. K. Rao (1969). A new estimation theory for sample surveys II. New Developments in Survey Sampling. Godambe and Sprott. New York, Wiley Inter-Science: 147-169.

Hartley, H. O. and R. L. Sielken (1975). "A "super-population viewpoint" for finite population sampling." Biometrics 31: 411-422.

Harville, D. A. (1977). "Maximum likelihood approaches to variance compnent estimation and to related problems." Journal of the Americnan Statistical Association 72(35): 320-340.

Harville, D. A. (1978). "Alternative formulations and procedures for the two-way mixed model." Biometrics 34: 441-453.

Henderson, C. R. (1984). Applications of linear models in animal breeding. Guelph, Canada, University of Guelph.

Henderson, C. R., O. Kempthorne, S. R. Searle and C. M. von Krosigk (1959). "The estimation of environmental and genetic trends from records subject to culling." Biometrics: 192-218.

Hinkelmann, K. and O. Kempthorne (1994). Design and analysis of experiments. Volume 1. Introduction to experimental design. New York, John Wiley and Sons.

Holt, D. (1989). Introduction to 'Disaggregated analysis: modelling structured populations'. Analysis of Complex Surveys. C. J. Skinner, Holt, D., and Smith,T.M.F. Chichester, Wiley: 209-220.

Horvitz, D. G. and D. J. Thompson (1952). "A generalization of sampling without replacement from a finite universe." Journal of the American Statistical Association: 663-685.

Institute, R. T. (2001). Sudaan: User Manual Release 8.1 Vol I and II. Research Triangle Park, North Carolina, Research Triangle Institute.

Johnson, N. L. and H. Smith (1969). New developments in Survey Sampling. New York, John Wiley and Sons.

Kempthorne, O. (1955). "The randomization theory of experimental inference." Journal of the Americian Statisticial Association 50: 946-967.

Kleffe, J. and J. N. K. Rao (1992). "Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model`." Journal of Multivariate Analysis 43: 1-15.

Konijn, H.S. (1962). “Regression analysis in sample surveys,” Journal of the American Statistical Association, 57:590-605.

Kovar, J. G., J. N. K. Rao and C. F. J. Wu (1988). "Bootstrap and other methods to measure errors in survey estimates." The Canadian Journal of Statistics 16: 25-45.

Laird, N. M. and J. H. Ware (1982). "Random-effects models for longitudinal data." Biometrics 38(4): 963-974.

Lencina, V., J. M. Singer and E. J. I. Stanek (2003). "Much ado about nothing: the mixed model controversy revisited." The American Statistician under review.

Lencina, V. B. (2002). Modelos de efeitos aleatorios e populacoes finitas. Statistics, USP. Sao Paulo, University of Sao Paulo.

Li, W. (2003). Use of random permutation model in rate standardization and calibration. Biostatistics and Epidemiology. Amherst, University of Massachusetts.

Li, W. and E. J. III. Stanek (2004). Covariance adjusted estimation under a design-based random permutation model, under review, Journal of Statistical Planning and Inference, 1/2004.

Liang, K.-Y. and S. L. Zeger (1995). "Inference based on estimating functions in the presence fo nuisance parameters." Statistical Science: 158-199.

Littell, R. and R. Wolfinger (1995). Mixed model analysis of data with the SAS system, Walt Disney World Dolphin - Salon V.

Liu, J. S. (2000). "MSEM dominance of estimators in two seemingly unrelated regressions." Journal of Statistical Planning and Inference 88: 255-266.

Lu, J. (2004). “Estimating parameters when considering the unobserved units as missing values in simple random sampling,” Master of Science Thesis, Department of Biostatistics and Epidemiology, UMASS, Amherst.

Lu, J., Stanek, E.J. III, and Puleo, E. (2004). “Estimating the population mean from a simple random sample when some responses are missing,” manuscript in preparation (c04ed02.doc).

McCarthy, P. J. and C. B. Snowden (1985). "The bootstrap and finite population sampling." Vital and Health Statistics Series 2 95(DHHS (PHS)85-1369).

McCulloch, C. E. and S. R. Searle (2001). Generalized, Linear, and Mixed Models. New York, John Wiley and Sons.

McLean, R. A., W. L. Sanders and W. W. Stroup (1991). "A unified approach to mixed linear models." The American Statistician 45(1): 54-64.

Merriam, P. A., I. S. Ockene, J. R. Hebert, M. C. Rosal and C. E. Matthews (1999). "Seasonal variation of blood cholesterol levels: study methodology." J Biol Rhythms 14(4): 330-9.

Morris, J. S. (2002). "The BLUPs are not 'best' when it comes to bootstrapping." Statistics and Probability Letters 56: 425-430.

Murray, D. M. (1998). Design and analysis of group-randomized trials. New York, Oxford University Press.

Neuhaus, J. M., J. D. Kalbfleisch and W. W. Hauck (1991). "A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data." International Statistical Review 59(1): 25-35.

Ockene, I. S., J. R. Hebert, J. K. Ockene, G. M. Saperia, R. Nicolosi, P. A. Merriam and T. G. Hurley (1996). Effect of physician-delivered nutrition counseling training and a structured office practice on diet and serum lipid measurement s in a hyperlipidemic population The Worcester-area trial for counseling in hyperlipidemia(WATCH). Archives of Internal Medicine.

Ockene, I. S., D. E. Chiriboga, E. J. I. Stanek, M. G. Harmatz, R. Nicolosi, G. Saperia, A. D. Well, P. A. Merriam, G. Reed, Y. Ma, C. E. Matthews and J. R. Hebert (2004). "Seasonal variation in serum cholesterol: treatment implications and possible mechanisms." Archives of Internal Medicine, in press.

Pfeffermann, D. and Nathan, G. (1981). “Regression analysis of data from a cluster sample,” Journal of the American Statistical Association. 76:681-689.

Pfeffermann, D. (1984). “On extensions of the Gauss-Markov Theorem to the case of stochastic regression coefficients,” Journal of the Royal Statistical Society B, 46:139-148.

Porter, R.M. (1973). “On the use of survey sample weights in the linear model,” Annals of Economic and Social Measurement, 2:141-158.

Prasad, N. G. N. and J. N. K. Rao (1990). "The estimation of the mean squared error of small-area estimators." Journal of the American Statistical Association 85: 163-171.

Prasad, N. G. N. and J. N. K. Rao (1999). "On robust small area estimation using a simple random effects model." Survey Methodology 25: 67-72.

Rao, J. N. K. (1999). "Some current trends in sample survey theory and methods." Sankhya:The Indian Journal of Statistics: Special issue on Sample Surverys 61(Series B, Pt.1): 1~57.

Rao, J. N. K. and C. F. J. Wu (1988). "Resampling interference with complex survey data." Journal of the American Statistical Association 83: 231-241.

Raudenbush, S. R. and A. S. Bryk (2002). Hierarchical linear models: applications and data analysis methods. London, Sage Publications.

Revankar, N. S. (1974). "Some finite sample results in the context of two seemingly unrelated regression equations." Journal of the American Statistical Association 69: 187-190.

Robinson, G. K. (1991). "That BLUP is a good thing: the estimation of random effects." Statistical Science 6(1): 15-51.

Royall, R. (1968). "An old approach to finite population sampling theory." American Statistical Association Journal: 1269-1279.

Royall, R. M. (1970). "Finite population sampling - on labels in estimation." The Annals of Mathematical Statistics 41(5): 1774-1779.

Royall, R. M. (1973). The prediction approach to finite population sampling theory: application to the hospital discharge survey. Rockville, Md, DHEW No. (HSM) 73-1329**:** 1-31.

Royall, R. M. (1976a). "Likelihood functions in finite population sampling theory." Biometrika 63: 605-614.

Royall, R. M. (1976b). "The linear least-squares prediction approach to two-stage sampling." Journal of the American Statistical Association 71: 657-664.

Royall, R. M. (1986). "The prediction approach to robust variance estimation in two-stage cluster sampling." Journal of the American Statistical Association 81: 119-123.

Royall, R. M. (1988). The prediction approach to sampling theory. Handbook of Statistics. P. R. Krishnaiah and C. R. Rao, Elsevier Science Publishers. Volume 6**:** 399-413.

Royall, R. M. (1992). "Robustness and optimal design under prediction models for finite populations." Survey Methodology 18: 179-185.

Royall, R. M. and W. G. Cumberland (1978). "Variance estimation in finite population sampling." Journal of the American Statistical Association 73: 351-358.

Royall, R. M. and J. Herson (1973a). "Robust estimation in finite populations. I." Journal of the American Statistical Association 68: 880-889.

Royall, R. M. and J. Herson (1973b). "Robust estimation in finite populations II: Stratification on a size variable." Journal of the American Statistical Association 68: 890-893.

Rust, K. (1985). "Variance estimation for complex estimators in sample survey." Journal of Official Statistics 1: 381-397.

Rust, K. and J. N. K. Rao (1996). "Variance estimation for complex surveys using replication techniques." Statistical Methods in Medical Research 5: 283-310.

Sarndal, C.-E. (1978). "Design-based and model-based inference in survey sampling." Scandinavian Journal of Statistics 5: 27-43.

Sarndal, C.-E., B. Swenson and e. al (1989). "The weighted residual technique for estimating the variance of the general regression estimator of the finite population total." Biometrika 76: 527-536.

Sarndal, C. E., B. Swensson and J. Wretman (1992). Model assisted survey sampling. New York, Springer-Verlag.

Sarndal, C.-E. and R. L. Wright (1984). "Cosmetic forms of estimators in survey sampling." Scandinavian Journal of Statistics 11: 146-156.

Scheffe, H. (1956a). "Alternative models for the analysis of variance." Annals of Mathematical Statistics 27: 251-271.

Scheffe, H. (1956b). "A 'mixed model' for the analysis of variance." Annals of Mathematical Statistics 27: 23-36.

Scheffe, H. (1959). The analysis of variance. New York, John Wiley and Sons.

Schensul, J. J., C. Huebner, M. Snow, R. Pino and L. Broomhall (2000). "The high, the money, and the fame: Smoking Bud among urban youth." Medical Anthropology Spring.

Schmidt, M., J. B. Erickson, P. S. Freedson, G. Markenson and L. Chasan-Taber (2002). "Physical activity patterns during pregnancy in a low income racially diverse population." American Jouranl of Epidemiology 155:S103.

Scott, A. and T. M. F. Smith (1969). "Estimation in multi-stage surveys." Journal of the American Statistical Association 64(327): 830-840.

Shao, J. (1996). "Resampling methods in sample surveys." Statistics and Probability Letters 27: 203-237.

Sitter, R. R. (1992a). "A resampling procedure for complex survey data." Journal of the American Statistical Association 87: 755-765.

Sitter, R. R. (1992b). "Comparing three bootstrap methods for survey data." The Canadian Journal of Statistics 20: 135-154.

Skinner, C. J. (1989). Chapter 2: Introduction to part A. Analysis of Complex Surveys. C. J. Skinner, D. Holt and T. M. F. Smith. New York, John Wiley and Sons: 23-58.

Srivastava, V. K. and D. E. A. Giles (1987). Seemingly unrelated regression equations models: estimation and inference. New York, Dekker.

Stanek, E. J. I., A. Well and I. Ockene (1999). "Why not routinely use best linear unbiased predictors (BLUPs) as estimates of cholesterol, per cent fat from Kcal and physical activity?" Statistics in Medicine 18: 2943-2959.

Stanek, E. J. I. (2002). Developing the Expected Value and Variance of an Expanded Set of Random Variables when N=3 and n=2, www-unix.oit.umass.edu/~cluster/ed/Results-unpub2002.html.

Stanek, E. J. I. and J. M. Singer (2003). Prediction of Random Effects in Unbalanced Cluster Sampling, http://www-unix.oit.umass.edu/~cluster/ed/outline/yr2002/C02ed43v1.pdf.

Stanek, E. J. I., J. M. Singer and V. B. Lencina (2004). "A unified approach to estimation and prediction under simple random sampling." Journal of Statistical Planning and Inference, 121:325-338.

Stanek, E. J. I. and J. M. Singer (2004). "Predicting random effects from finite population clustered samples with response error." JASA in press.

Stanek, E.J. III, Singer, J.M., Li, W., and Reed, G. (2004). "When BLUPs are bad," presented at the ENAR Biometrics Meetings, March 31, Pittsburg, Pa.

Stukel, D. M., M. A. Hidiroglou and C.-E. Sarndal (1996). "Variance estimation for calibration estimators: a comparison of Jackknifing versus Taylor Linearization." Survey Methodology 22(2): 117-125.

Stukel, D. M. and J. N. K. Rao (1997). "Estimation of regression models with nested error structure and unequal error variances under two and three stage cluster sampling." Statistics \& Probability Letters 35: 401-407.

Tang, B. (1999). "Balanced bootstrap in sample surveys and its relationship with balanced repeated replication." Journal of Statistical Planning and Inference 81: 121-127.

Thompson, M. E. (1997). Theory of Sample Surveys. London, Chapman&Hall.

Valliant, R. (2002). "Variance estimation for the general regression estimator." Survey Methodology 28: 103-114.

Valliant, R., H. A. Dorfman and R. M. Royall (2000). Finite population sampling and inference. New York, John Wiley and Sons.

Verbeke, G. and G. Molenberghs (2000). Linear mixed models for longitudinal data, Springer, New York.

Wolter, K. M. (1985). Introduction to variance estimation. New York, Springer-Verlag.

Wu, C. and R. R. Sitter (2001). "A Model-calibration Approach to Using Complete Auxiliary Information From Survey Data." Journal of the American Statistical Association(Theory and Methods) 96(453): 185-193.

Yeo, D., H. Mantel and T.-P. Liu (1999). "Bootstrap variance estimation for the National Population Health Survey." ASA Proceedings of the Section on Survey Research Methods: 778-783.

Zellner, A. (1962). "An efficient method of estimating seemingly unrelated regression and tests for aggregation bias." Journal of the American Statistical Association 57: 348-368.

Zellner, A. (1963). "Estimators of seemingly unrelated regression equations: some exact finite sample results." Journal of the American Statistical Association 58: 977-992.