ABSTRACT

Title of Document: RANDOMIZATION-BASED INFERENCE ABOUT LATENT VARIABLES FROM COMPLEX SAMPLES: THE CASE OF TWO-STAGE SAMPLING

Tiandong Li, Doctor of Philosophy, 2012

Directed By: Professor Robert J. Mislevy
Department of Measurement, Statistics and Evaluation

In large-scale assessments, such as the National Assessment of Educational Progress (NAEP), plausible values based on Multiple Imputations (MI) have been used to estimate population characteristics for latent constructs under complex sample designs. Mislevy (1991) derived a closed-form analytic solution for a fixed-effect model in creating plausible values assuming a classical test theory model and a stratified student sample and proposed an analogous solution for a random-effects model to be applied with a two-stage student sample design. The research reported here extends the discussion of this random-effects model under the classical test theory framework. Under the simplified assumption of known population parameters, analytical solutions are provided for multiple imputations in the case of the classical test theory measurement model and two-stage sampling and their properties are verified in reconstructing population properties for the unobservable latent variables. With the more practical assumptions of unknown population and cluster means, this study empirically examines the reconstruction of population attributes. Next, properties of sample statistics are examined. Specifically, this research explores the impact of the variance components and sample sizes on the sampling variance of the MI-based estimate for the population mean. Findings include significant predictors and influential factors. Last, the relationships between the sampling variance of the estimate of the population mean based on the imputations and those based on observations of the true score and the observed score are discussed. The sampling variance based on the imputed score is expected to be the higher boundary of that based on the observed score, which is expected to be the higher boundary of that based on the true score.

RANDOMIZATION-BASED INFERENCE ABOUT LATENT VARIABLES FROM COMPLEX SAMPLES: THE CASE OF TWO-STAGE SAMPLING

By

Tiandong Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2012

Advisory Committee:
Professor Robert J. Mislevy, Chair
Dr. Frank F. Jenkins
Professor Partha Lahiri
Professor George B. Macready
Research Professor Keith F. Rust

DEDICATION

This dissertation is dedicated
to my parents, Zuntang Li and Lianbin Zhang,
to my wife, Lanlan Yin, and
to my son to be born, Alan Li.

ACKNOWLEDGEMENTS

First and foremost, I wish to convey my deepest gratitude to my dissertation advisor and mentor, Dr. Robert Mislevy, who showed me the way to become a researcher and set an excellent example for me. His patience, encouragement, and enthusiasm made this dissertation possible.

I would also like to thank Dr. Frank F. Jenkins, Dr. Partha Lahiri, Dr. George Macready, and Dr. Keith Rust for their guidance and insightful suggestions.

I would not have gotten this far without the continued support of my employer, Westat, and specifically the support of the director of the statistical group David Morganstein. Thank you for providing me flexibility at work that allowed me to accommodate my study with my full time job. Also, I would like to thank my colleagues and friends who provided valuable feedbacks to my research: Dr. Robert Fay, Dr. Graham Kalton, Dr. Konrad Noben-Trauth and Dr. Judith Strenio.

To my parents Zuntang Li and Lianbin Zhang, your unconditional love and continuous encouragement supported me throughout this long journey of dissertation work. Thank you for your faith in me through all these years. I hope that what I have accomplished makes you happy and proud.

Most importantly, thank you Lanlan for everything you have done to ensure my achievements. Your continuous understanding, sacrifice and love made this dissertation possible. I thank your parents (whom I consider my parents as well) for their selfless contribution to our family.

Finally, I dedicate this thesis to our coming son, Alan, as my first gift in his life.

Table of contents

List of Tables

List of Figures

Chapter 1 : Introduction

1.1 Background

In psychometrics, measurement models provide a platform for explaining theoretical latent constructs underlying observed item responses. Traditional measurement models are generally developed under the assumption of simple random sampling (SRS) of individuals, as the structures of interest concern (often complex) within-person patterns of response. However, in large-scale educational assessments, such as the National Assessment of Educational Progress (NAEP), data are collected under complex sample designs, which include the following three components: unequal weights, stratification, and clustering (Rust, Krenzke, Qian, & Johnson, 2001) . In addition, to reduce the test burden on respondents, students take only a subset of the test items. The subsets are selected using the multiple-matrix item sampling method. This item sampling design is handled by applying an Item Response Theory (IRT) latent-variable model to estimate student proficiency or ability, and the analyst calculates and reports results on the scale of the latent variable. However, an efficient IRT estimator of individuals' proficiency can be seriously biased in estimating the population distribution of the proficiency scores (Lord, 1959; Mislevy, Beaton, Kaplan, & Sheehan, 1992). To avoid the estimation of individual student latent-variable parameters when estimating population characteristics, population characteristics can be calculated based on their conditional expectation in marginal analyses (Mislevy, 1984). In addition, this approach can jointly handle the latent variable model and complex student sampling.

Because the closed-form solution for the conditional expectation is only available for special cases, an alternative called plausible values, based on Rubin's

(1987). Multiple Imputation (MI), has been used to allow secondary researchers to estimate latent trait distributions in large-scale educational assessments. Although the methods have proven useful, they can be difficult to understand in applications involving both complex measurement models and sampling designs. To provide intuition, Mislevy (1991) derived a closed-form analytic solution for a fixed-effect model for MI assuming a classical test theory model and a stratified student sample. The results provide insight into the elements and properties of the procedure, and ground intuition for more complex applications. In the same article Mislevy proposed an analogous solution for a random-effects model to be applied with a two-stage student sample design. However, no proof or further discussion was provided on the character of the solution, nor has one appeared in the subsequent literature. Research presented here fills in this gap, providing analytic derivations for the necessary components of the solution and demonstrating its properties in a range of circumstances with simulated data. As such, we provide additional conceptual grounding for practitioners who develop and/or use plausible values. The study design of NAEP is discussed as a representative example in this study.

1.2 Research Purposes and Questions

Here, we derive formulas for multiple imputations in the case of the classical test theory (CTT) measurement model and two-stage sampling, verify their properties in reconstructing population properties for the unobservable latent variables, or θs, and empirically examine the reconstruction of population attributes and the properties of sample estimates with the more practical assumptions of unknown population and cluster means.

Specifically, the research consists of two parts:

Under the simplified assumptions of known population parameters, analytic demonstration is provided for the construction of multiple imputations of θ and derivations of desired properties of the imputations for the case of two-stage cluster sample design.

Under the relaxed assumption of unknown population and cluster means, simulation-based demonstration is provided for the construction of multiple imputations of θ and exploration of properties of the imputations for the case of two-stage cluster sample design.

1.3 Organization of the Chapters

This study is presented in six chapters.

Chapter two gives a review of the relevant literature. The first five sections briefly review aspects - test theory, multiple matrix sampling, clustered population and random-effect model, complex survey sampling, and randomization-based inference in survey sampling - provide the basis of the research framework for this study, which is illustrated in section 2.6 - Multiple Imputation for latent variables in complex sample surveys.

Chapter three begins the first part of the research results with an analytic discussion of the Multiple Imputation approach for latent variables in two-stage samples. It derives the general form of the posterior distribution of MI and the specific case of classical test theory.

Chapter four gives the analytical solution when the population parameters are known to show the reproduction of population characteristics within the MI dataset structure.

Chapter five presents the second part of the research, the simulation study of the situation where the population and cluster means are not known. It consists of five subsections. The first discusses the three major research questions the simulation study is designed to explore. The second section explains the construction of imputations for the case of unknown means. Next, the study method and data generation process, are described. The fourth section analyzes the simulation results, and the fifth section presents the analysis results in terms of the three research questions posed at the outset of the simulation study.

Chapter six discusses the importance of this study, summarizes the major findings, and addresses the limitations of the study, concluding with some suggestions for future research.

Chapter 2 : Literature Review

2.1 Test Theory

Educational test theory provides statistical and methodological tools to make inference about examinees' knowledge, skills, and accomplishments. Since the first text on test theory published by E. L. Thorndike in 1904, researchers have extended test theory from Classical Test Theory (CTT) to generalizability theory, item response theory (IRT) and the analysis of relationships among scores from different tests, including factor analysis, structural equations modeling, and multitrait-multimethod analysis (Mislevy, 1996). This research uses a straightforward measurement model, classical test theory (CTT), which yields closed-form solutions that support intuition for more complicated measurement models such as IRT.

2.1.1 Classical Test Theory

The foundation of CTT was laid by Spearman (1904a, 1904b, 1907, 1913). This model was extensively presented by Gulliksen (1950) and developed more rigorously by Lord and Novick (1968). As shown in Crocker and Algina (1986), the CTT model envisions an observed test score as the composite of two hypothetical components – a true score and a random error component – expressed in the form

$$X_i = \theta_i + E_i \tag{2.1}$$

where X_i represents the observed test score of the *i*th examinee; θ_i, the individual's true score; and E_i, a random error component. Both X_i and E_i are random variables in terms of repeated observations for examinee *i*, and θ_i is a constant for examinee *i*. The assumptions of the CTT models are as below:

1) The mean of the random error is zero $(E(E_i)=0)$.

2) The correlation between true and error scores of a test for a population of examinees is zero $(\rho_{\theta E}=0)$

3) The correlation between error scores from two parallel tests is zero $(\rho_{E_1 E_2}=0)$

Under assumption 2, the relationship of the variances of the three components in the CTT model can be shown to be

$$\sigma_X^2 = \sigma_\theta^2 + \sigma_E^2 \tag{2.2}$$

The reliability coefficient defined as the ratio of true score variance to observed score variance can be expressed as

$$\rho_{X_1 X_2} = \frac{\sigma_\theta^2}{\sigma_X^2} \tag{2.3}$$

which shows the proportion of observed score variance explained by the true score variance. In CTT, scores are obtained over a large number of items and are treated as continuous. The reliability coefficient can be approximated by the estimates of the internal reliability across items (e.g., Cronbach’s alpha coefficient).

CTT is a longstanding, satisfactory method used in the area of standardized testing. An advantage with CTT is that the model relies on weak assumptions that are easy to meet by standardized testing procedures. In addition, with its linear structure and the additional assumption of normally distributed errors, the CTT model is relatively simple and easy to interpret. We use this model in this study for simplicity and the intuition that the results provide.

2.1.2 Item Response Theory

Item Response Theory is essentially a mathematical model for the probability of a correct response to an item, given the person’s proficiency parameter and one or more parameters for each item (Mislevy, 1989). Both the person's proficiency and item difficulties are positioned on the same latent scales. A major advantage of IRT over CTT is the proficiency invariance interpretation with respect to selection of items. That is, the expected student proficiency score is independent of the set of items administered to him or her. This feature of the model allows IRT to handle the Multiple Matrix Sample described in the next section. Although this study focuses on a simpler test theory, IRT is the model that has actually been implemented in large-scale assessments including NAEP and hence motivated the choice of exercising the Multiple Imputation discussed in section 2.6.

2.2 Multiple Matrix Sampling

Along with survey sampling of students, Multiple Matrix Sampling of test items is widely employed in educational assessment (Educational Testing Service & National Center for Education Statistics, 1999). In Multiple Matrix Sampling, random subsamples of students are administered subsets of the entire pool of assessment items. This design permits a satisfactory precision level in estimating population characteristics and a complete coverage of the assessment framework while minimizing the time burden for each student. Researchers have shown that population characteristics can be estimated accurately without precise measurement of individual students (Lord, 1962; Lord et al., 1968; Sirotnik & Wellington, 1977). In fact, the population item-score mean is estimated most efficiently when each student in the group is assigned one distinct item from each objective reporting area. Therefore, a highly detailed curricular evaluation with 30-50 objectives can be implemented by administering a test form of even fewer than 30 items for each student, as long as the students who receive items from a given objective are a representative sample. The length of the test is still within a reasonable limit. These findings in the 1970s led to the use of multiple matrix sampling designs in educational assessment for efficiently estimating distributions of performance in the population (or subpopulations) in large-scale assessments such as NAEP.

The type of matrix sampling used by NAEP is called focused, balanced incomplete block (BIB) spiraling. The "focused" part of NAEP's matrix sampling method requires each student to answer questions from only one subject area. The "BIB" part of the method ensures that students receive different interlocking sections of the assessment forms, enabling NAEP to check for any unusual interactions that may occur between different samples of students and different sets of assessment questions. "Spiraling" refers to the method by which test booklets are assigned to pupils, which ensures that any group of students will be assessed using approximately equal numbers of the different versions of the booklet (Educational Testing Service & National Center for Education Statistics, 1999). Because of BIB spiraling, NAEP can sample enough students to obtain precise results for each question while consuming an average of about an hour and a half of each student's time.

The original NAEP surveys in the 1970s focused on item-level results. Beginning in the assessment of 1984, however, it was desired to produce distributions of proficiency in populations and subpopulations of students. An IRT model is desirable in estimating student proficiency based on data from multiple matrix item sampling. The number of items arranged for each student is too small to make an accurate estimate of the proficiency, which typically ranges between 5 and 15 items in a given reporting area. However population characteristics are estimated on IRT scales directly from survey responses through marginal estimation procedures, as discussed in section 2.6.1. The plausible values provided on public use data sets allow secondary analysts to reproduce the official estimates and to carry out analyses of their choice on the NAEP IRT scales.

2.3 Clustered Population and Random-Effect Model

The population of students in educational assessments is clustered within naturally occurring organizational units, such as classes, schools, and districts. This study is concerned with the population parameters in a two-level clustered population. The traditional population in statistical studies assumes independence of observations. However, when students are clustered within natural units, the responses from the same cluster are correlated with each other in some degree. For example, students in one school may tend to achieve higher assessment scores than students in another school in general. Therefore, the scores of students are not independent to each other. Multilevel modeling allows researchers to model this nonindependence and views the population structure as of potential interest (Goldstein, 2010). While this study only deals with a simple case of multilevel modeling, the so-called random-effects model and mixed-effects model (Elston & Grizzle, 1962) or the random-intercept model (Raudenbush & Bryk, 2002), interest in applying more complex multilevel modeling in large-scale assessments has been increasing, e.g. Braun, Jenkins, & Grigg (2006). In the simple two-level model in this study, independence will be assumed at the cluster level and within each cluster. The assumptions made about the variances and covariances are stronger than a traditional analysis.

This study considers a random-effects model with equal cluster size in the clustered population. The population variance structure is assumed to consist of between-cluster variance and within-cluster variance. Measurement error will add a third level of variance within students.

If clusters are indexed by *k* and subjects within a cluster are indexed by *i*, the observed score X_{ik} in a CTT can be rewritten as:

$$\begin{aligned} X_{ik} &= \theta_{ik} + E_{ik} \\ &= \nu_k + \gamma_{ik} + E_{ik} \end{aligned} \tag{2.4}$$

where θ_{ik} is the true score of the *i*th examinee within the *k*th cluster, E_{ik} is the measurement error of an individual person within cluster *k*, and γ_{ik} is the deviation of the individual's true score from the cluster mean ν_k. The variance of true score θ_{ik} in the population can be expressed in two components: between-cluster variance σ_b^2 and within-cluster variance σ_w^2, that is $\sigma_\theta^2 = \sigma_b^2 + \sigma_w^2$. Thus, variance of observed score X_{ik} in the population can be expressed in three components:

$$\sigma_X^2 = \sigma_b^2 + \sigma_w^2 + \sigma_e^2 \tag{2.5}$$

The random-effects model that Mislevy (1991) proposed shows the distribution of the latent variable as follows:

$$\nu_k \sim N(\mu, \sigma_b^2) \tag{2.6}$$

$$\text{and } \theta_{ik} \mid (z = k) \sim N(\nu_k, \sigma_w^2) \tag{2.7}$$

where μ is the overall population mean of the latent variable θ_{ik}, ν_k is the cluster mean when the cluster index *z* equals *k*, and σ_b^2 and σ_w^2 represent between-cluster variance and within-cluster variance for the population. Hence, the distribution of γ_{ik} is as follows:

$$\gamma_{ik} \mid (z = k) \sim N(0, \sigma_w^2) \tag{2.8}$$

These models show the mechanism for how student scores are modeled. As stated in the research purposes, this study explores the properties of multiple imputations (aka plausible values) in terms of reproducing the population statistics of the true score shown above, which include population mean, cluster means and the variance components.

2.4 Complex Survey Sampling

As Bock (1982) indicated, survey sampling designs gained popularity in efficiently collecting social information during the 1960s. They had already been employed to collect information about educational aspirations and attitudes.
When properly undertaken, a sample survey provides an objective, efficient, and valid method of obtaining the characteristics of an entire population from only a small part of that population (Frankel & Frankel, 1987). Complex sample designs feature at least one of three components: unequal probability of selection, stratification, and clustering e.g. Cochran (1977). These designs are usually motivated by cost constraints and administrative reasons, as well as estimation accuracy for the population or sub-population.

In the naturally clustered population in educational assessments, such as students in schools, treating schools as the first-level sampling unit in a (multi-stage) cluster sample of students saves administrative costs and traveling expenses by not going to a large number of schools which may only have a few sampled students each. Although a larger number of students will be needed to gain the same accuracy level as from a Simple Random Sample, a cluster design will reduce the number of schools one has to visit and therefore probably reduces the cost of data collection. As an example of multi-stage probability sampling design, the sample for the NAEP 1998 national assessment was drawn via four stages of selection (Rust et al., 2001), treating geographic areas and schools, etc. as clusters: 1) the selection of Counties or groups of counties, Primary Sampling Units (PSUs); 2) the selection of elementary and secondary schools within PSUs; 3) the assignment of sessions by type and of sample types to sampled schools; and 4) selection of students within schools and their assignment to session types.

2.5 Randomization-based Inference in Survey Sampling

As pointed out by Cassel, Sarndal and Wretman (1977), two competing basic philosophies in the theory of inference for finite populations are design-based inference and model-based inference. The design-based inference sees the primary source of randomness is the probability ascribed by the sampling design to the various subsets of the finite population $\{1, ..., N\}$. On the contrary, in model-based theory of inference in survey sampling, the values $(y_1, \cdots, y_N)$ associated with the N units of the population units are viewed as the realized outcome of random variables $(Y_1, \cdots, Y_N)$ having an N-dimensional joint distribution.

Randomization-based inference is used for most of the work in this study. It is the traditional and dominating mode of inference in survey sampling, following the milestones of literature starting with Neyman (1934) and subsequent work including Hansen, Hurwitz and Madow (1993), Mahalanobis (1946), Kish (1965) and Cochran (1977), etc. As discussed in Kalton (1983), with the large samples typical of most surveys, survey practitioners are reluctant to use model-based estimators of descriptive parameters because of the potential estimation bias resulting from any misspecifications of the model. However, elements of the model-based approach are required to implement Rubin's multiple imputation scheme for latent variables.

For a finite population with N units, indexed by i, the values of a survey item can be denoted as $\mathbf{Y} = (y_1, y_1, \ldots, y_N)$. To conduct randomization-based inference, y_i 's are treated as fixed but unknown values. The statistics of interest can be represented as $S \equiv S(\boldsymbol{Y}, \boldsymbol{Z})$, where $\boldsymbol{Z}$ is the vector of design variables, which are known for all units before observation. Let IND = $(I_1, I_2, \ldots, I_N)$ represent the sample indicator, a vector of random variables, where I_i = 1 if unit i in the population is in the sample and I_i = 0 if unit i is not in the sample. According to a sample design-based on the probability of IND, an unbiased sample statistics $s \equiv s(Y_{sample}, Z, IND)$ and an estimator of sampling variance $U \equiv U(Y_{sample}, Z, IND)$ can usually be constructed. Clustering in the sample design can be reflected by the linked probability of I_i for units in the sample cluster. Inferences from sample statistics s to the population statistics S are based on the distribution of s in repeated samples of Y_{sample} under an identical sample design. Randomization-based inferences are then based on the normal distribution from large-sample approximations: $\frac{s-S}{\sqrt{U}} \sim N(0,1)$.

The pieces of the theories reviewed in sections 2.1-2.5 provide the basis of the research framework for this study to be illustrated in section 2.6.

2.6 Multiple Imputation for Latent Variables in Complex Sample Surveys

Mislevy (1991) illustrated the theoretical framework for the estimation of distributions of latent variables in finite populations, when the sample is drawn under a complex sampling design. Latent variables in a sample survey will be treated as survey variables with missing values for all respondents. By the nature of the latent variable model, the assumption of missing at random (MAR) is satisfied. Knowledge about the latent variable θ can be fully reflected by a posterior distribution given the observed data, namely the design (sampling frame) variable $\boldsymbol{Z}$, background survey variables Y, and item responses X. To estimate a scalar $S \equiv S(\boldsymbol{\Theta}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, a certain function of these four types of variables, three building blocks are needed.

2.6.1 The Sampling Model

The first building block is called the sampling model, which makes a randomization based inference about the population characteristics S from the sample statistics $s \equiv s(\theta, X, Y, \boldsymbol{Z}, IND)$. When θ is known, the traditional randomization based inference in sampling statistics relies on the central limit theorem. When the sample size is large, the sample statistics, such as the sample mean, have the following distribution:

$$\frac{s - S}{\sqrt{Var(s)}} \sim N(0,1) \qquad (2.9)$$

As s cannot be calculated when we don't know θ, the conditional expectation may possibly be calculated instead, based on the predictive (or posterior) distribution of the latent variable θ (Rubin, 1977):

$$E_{\theta}(s|X, Y, \boldsymbol{Z}) = \int s(\theta, X, Y, \boldsymbol{Z}, IND) p(\theta|X, Y, \boldsymbol{Z}) d\theta \qquad (2.10)$$

where all variables are fixed while $\boldsymbol{Z}$ is known and the value of X and Y will become known for sampled units based on a sample design. This approach makes it possible to estimate population characteristics of the latent variables, such as means and proportions of students above specified proficiency levels, directly from the observed responses, avoiding the steps of calculating scores for individual students.

To obtain the predictive distribution $p(\theta|X, Y, \boldsymbol{Z})$ using Bayes Theorem, the other two building blocks are needed, the population model and the latent variable model.

2.6.2 The Population Model

The population model assumes that the distribution of θ, given the survey collateral variable Y and the design variable $\boldsymbol{Z}$, is of the form $p(\theta|Y, \boldsymbol{Z}, \alpha)$, where the unknown parameter of the distribution is represented by α.

For example, Mislevy (1991) thoroughly discussed the fixed effect model in the presence of collateral survey variables. (In modeling practice the stratification design variables can be treated as collateral survey variables.) The conditional distribution for the fixed effect population model is defined as

$$\theta|Y \sim N\left(\delta' Y, \sigma^2_{\theta|Y}\right) \tag{2.11}$$

where δ represents the regression parameters of Y on θ and

$$\sigma^2_{\theta|Y} \equiv Var(\theta|Y) = 1 - R^2 \tag{2.12}$$

with R^2 showing the proportion of variance of θ explained by Y.

In this study, the population model follows the distribution of the two-level clustered population discussed in section 2.3. Its parameters (α) are the population mean, the cluster means, and variance at each level.

2.6.3 The Latent Variable Model

The latent variable model assumes that the distribution of the item response X, given the latent variable θ, is of the form $p(X|\theta, \beta)$, where the unknown parameter of the distribution is represented by β.

As discussed in section 2.1.1, this study uses a CTT model as the latent variable model. The unknown parameter is the variance of the error term.

Using Bayes theorem, the posterior distribution of θ, $p(\theta|X, Y, \boldsymbol{Z})$, can be expressed as a function of the population model and the latent variable model, following Mislevy (1991):

$$p(\theta|X,Y,\boldsymbol{Z},\alpha,\beta) = K_{\alpha\beta}p(X|\theta,Y,\boldsymbol{Z},\alpha,\beta)p(\theta|Y,\boldsymbol{Z},\alpha,\beta) \tag{2.13}$$

where the constant $K_{\alpha\beta} = 1/p(X|Y,\boldsymbol{Z},\alpha,\beta)$ depends on α and β, but not θ. That is, the posterior distribution can be derived from the normalized product of 1) the likelihood function of θ, which is the conditional probability of X given θ, from the latent variable model and 2) the conditional distribution for θ given the background and design variables, from the population model. That is,

$$\begin{aligned} p(\theta|X,Y,\boldsymbol{Z},\alpha,\beta) &= K_{\alpha\beta}p(\mathrm{X}|\theta,Y,\boldsymbol{Z},\alpha,\beta)p(\theta|Y,\boldsymbol{Z},\alpha,\beta) \\ &= K_{\alpha\beta}p(\mathrm{X}|\theta,\beta)p(\theta|Y,\boldsymbol{Z},\alpha) \end{aligned} \tag{2.14}$$

Under the fixed effect model, given the same CTT latent variable model and the fixed effect population model, the posterior distribution is

$$\theta|X,Y{\sim}N\left(\bar{\theta},\sigma^2_{\theta|XY}\right) \tag{2.15}$$

where, as in Kelley (1923), $\bar{\theta} \equiv E(\theta|X,Y) = \rho_c X + (1-\rho_c)\ \delta' Y$ and $\sigma^2_{\theta|XY} \equiv Var(\theta|X,Y) = (1-\rho_c)\sigma^2_{\theta|Y} = (1-\rho_c)(1-R^2)$, with ρ_c, the "conditional reliability" of X given Y, as $\rho_c \equiv \frac{\sigma^2_{\theta|Y}}{\sigma^2_{\theta|Y}+\sigma^2_e}$.

2.6.4 Assumption for Imputation - Missing at Random

Under the framework of MI, to estimate characteristics of latent variables in a sample survey, the latent variables are treated as survey variables with missing values for all respondents.

Most MI methods require the assumption of missing at random (MAR) (Rubin, 1977). That is, the probability that the observation is missing does not depend on the value of the missing observation, given the values of the observed values and the value of any background variables. When treating latent variables as missing, this assumption holds by nature, as the latent variables are missing no matter what their values are. Thus, all knowledge about subjects' latent variables are conveyed by the predictive distribution $p(\theta|X, Y, Z, \alpha, \beta)$, upon which the imputation will be based.

2.6.5 The construction of Multiple Imputations for Latent Variables

Based on the framework described in the previous section, the population characteristics are estimated using the conditional expectation in the sampling model. However, closed-form solutions for the integral equations can only be calculated for special cases (Mislevy, Johnson, & Muraki, 1992). For example, the closed form of the posterior distribution is not available for data analysis when the latent variable model is an IRT measurement model, which is used in NAEP and other educational assessments. As an alternative method, stochastic, or Monte Carlo, integration based on random draws from posterior distributions of each sampled student is employed in estimating the conditional expectation in educational assessments.

Although the development of Markov chain Monte Carlo (MCMC) methods seem to overcome the computational difficulties described above to a large extent (Rao, 2003), we choose to develop the posterior distribution in a closed form in a case in which this is possible, to add insight to the MI process.

Also known as Multiple Imputations, random draws from posterior distributions are carried out several times to form sets of "plausible values." Additionally, MI or the plausible values provides "complete" data sets that the standard statistical methods can be applied to by secondary researchers. With the multiply-imputed data sets, each of the imputed complete data sets is analyzed by standard methods—including randomization-based estimates of population statistics and accompanying sampling variances. Inferences about statistics of interest will be made based on the combination of estimates of within-imputation and between-imputation variances.

Specifically, Rubin's estimates for a statistic and its sampling variance calculated using MI is carried out as follows, for the latent variable situation modeled as$K_{\alpha\beta}p(\mathrm{X}|\theta,\beta)p(\theta|Y,\boldsymbol{Z},\alpha)$.

1. Estimate the posterior distribution of the parameters β of the latent variable model and α of the population model, or $p(\alpha,\beta|X,Y,\boldsymbol{Z})$.
2. Create *M* imputed datasets $(\theta_{(m)},X,Y)$.
    a. Randomly draw a value $(\alpha,\beta)_{(m)}$ for the *m*-th data set from $p(\alpha,\beta|X,Y,\boldsymbol{Z})$.
    b. For each respondent *i* in the *m*-th data set, draw a value from the predictive distribution $p(\theta|x_i,y_i,\boldsymbol{z_i},(\alpha,\beta)=(\alpha,\beta)_{(m)})$. The resulting $\theta_{i(m)}$ values are the imputed values.
3. Using the multiple imputed data, calculate the point estimate and variance of the statistics *S*.

Rubin's formulation of the variance of a statistic based on m pseudo datasets starts with the within-imputation sampling distribution of the statistic, $N(s_{(m)},U_{(m)})$, where $s_{(m)}$ is the point estimate of some statistic of interest calculated on imputation set m and $U_{(m)}$ is the estimate of sampling variance, treating the imputations as if there were known true values. The following statistics can be calculated: mean of the estimates $s_{(m)}$ and within imputation estimate of sampling variance $U_{(m)}$, as averages over pseudo data sets, between imputation variance B_M , and total variance V_M , where

$$s_M = \frac{\sum s_{(m)}}{M}$$

$$U_M = \frac{\sum U_{(m)}}{M} \qquad (2.16)$$

$$B_M = \frac{\sum (s_{(m)} - s_M)^2}{(M-1)}$$

and $V_M = U_M + (1 + 1/M) B_M$

The development of the form of the imputation for this study is provided in chapter 3.

Chapter 3 : Multiple Imputation Approach for Latent Variables in Two-Stage Samples

Mislevy (1991) proposed the multiple imputation method along with the two-stage random-effect model, which was suggested to reproduce population variance components of the true score. As no detailed discussion was provided about how this imputation was formed, this chapter discusses this issue explicitly.

3.1 General Form

For the two-stage sample, the population model can be written as two levels:

1) The cluster level model:

$$\nu_k|(\mu, \sigma_b^2) \sim N(0, \sigma_b^2) \tag{3.1}$$

2) The examinee level model:

$$\theta_{ik}|(\nu_k, \sigma_w^2) \sim N(\nu_k, \sigma_w^2) \tag{3.2}$$

For a given form of the latent variable model $p(\mathrm{X}|\theta, Y, \boldsymbol{Z}, \alpha, \beta)$, we can construct the posterior distribution $p(\theta|X, Y, \boldsymbol{Z}, \alpha, \beta)$, from which the multiple imputations will be drawn.

3.2 The Case of Classical Test Theory

As this study employs the CTT model and the clustered population, the posterior distribution of θ_{ik} can be built in two stages: 1) the posterior distribution of the cluster mean of the true score conditioning on the individual observed scores in cluster *k* and higher level parameters, including μ, σ_b^2, σ_w^2 and σ_e^2; and 2) the posterior distribution of the true score of individual person conditioning on the cluster mean, all individual observed data and higher level parameters including ν_k, σ_w^2 and σ_e^2.

Given the latent variable model expressed as below

$$X_{ik}|(\theta_{ik},\sigma_e^2) \sim N(\theta_{ik},\sigma_e^2) \tag{3.3}$$

$$\text{and } \bar{X}_k|(\nu_k,\sigma_w^2,\sigma_e^2) \sim N\left(\nu_k,\frac{\sigma_w^2+\sigma_e^2}{I}\right) \tag{3.4}$$

where *I* is the sample size of the cluster, and the population model shown in section 3.1, normal posterior distributions can be derived. Within clusters, the posterior distribution is

$$p(\theta_{ik}|X_{ik},\nu_k,\sigma_w^2) \propto \; p(X_{ik}|\theta_{ik},\sigma_e^2)p(\theta_{ik}|\nu_k,\sigma_w^2) \tag{3.5}$$

and between clusters

$$p(\nu_k|\bar{X}_k,\mu,\sigma_b^2,\sigma_w^2,\sigma_e^2) \propto p(\bar{X}_k|\nu_k,\sigma_w^2,\sigma_e^2)p(\nu_k|\mu,\sigma_b^2) \tag{3.6}$$

Given the normal assumption of the population model and the latent variable model at both stages, the posterior distributions are resolved to be, for the true score of individual person within clusters,

$$\theta_{ik}|(X_{ik},\nu_k,\sigma_w^2) \sim N(\rho x_{ik}+(1-\rho)\nu_k,(1-\rho)\sigma_w^2) \tag{3.7}$$

where $\rho=\frac{\sigma_w^2}{\sigma_w^2+\sigma_e^2}$ is the within-cluster, examinee-level reliability coefficient; and for the cluster mean of the true score,

$$\nu_k|(X_{1k}\ldots X_{Ik},\mu,\sigma_b^2,\sigma_w^2,\sigma_e^2) \sim N(\lambda\bar{x}_k+(1-\lambda)\mu,(1-\lambda)\sigma_b^2) \tag{3.8}$$

where $\lambda=\frac{\sigma_b^2}{\sigma_b^2+(\sigma_w^2+\sigma_e^2)/I}$ is the cluster-level reliability coefficient and I is the number of subjects in a cluster.

Basically, the posterior mean at both the individual level and the cluster level is a weighted average of the population mean and the mean of the appropriate observed data.

An imputation for the cluster mean ν_k is $\lambda\bar{x}_k + (1-\lambda)\mu + g_k$, where g_k is a random draw from $N(0,(1-\lambda)\sigma_b^2)$ and an imputation for the latent variable θ_{ik} is

$$\rho x_{ik} + (1-\rho)[\lambda\bar{x}_k + (1-\lambda)\mu + g_k] + f_{ik} \qquad (3.9)$$

where f_{ik} is drawn from $N(0,(1-\rho)\sigma_w^2)$. Random terms g_k and f_{ik} are drawn from normal distributions with variances equal to the posterior variances of ν_k and θ_{ik}, respectively. By adding these two terms, the variances of the imputations for cluster means and for individual scores are unbiased. These two terms are referred to as variance reconstruction terms in the rest of the thesis. The next chapter derives formulas to show the unbiasedness of estimates based on the imputation. That is, the expected values of imputations so constructed, when population parameters are known, correctly reproduce the population latent-variable characteristics.

Chapter 4 : Analytical Solution with Known Population Parameters

At the first stage of work, we construct imputations with the higher level parameters μ and σ^2's treated as known, in order to demonstrate the reproduction of population characteristics within the MI dataset structure.

Mislevy (1991) demonstrated that the use of either maximum likelihood or Bayesian estimates for individuals' θs produced biased results for population variance components. The same paper proposed an approach to generating multiple imputations in the two-stage random-effects model, which were suggested to reproduce variance components, but no proof has ever been shown in the literature. The research in this dissertation provides results for the random-effects model that are analogous to Mislevy's analysis results for the fixed effects model.

This chapter shows that the within cluster estimator $\tilde{\nu}_k$ and population estimator $\tilde{\theta}_{ik}$ used in the imputation are unbiased when the population parameters μ, σ_b^2, σ_w^2 and σ_e^2 are known.

The posterior distribution of θ given the observed scores and known parameters is derived analytically. The resulting equations illustrate the desirable properties of the imputations for the case of a two-stage cluster sample design. The result can provide intuitive understandings for more complex cases.

4.1 Derivation of Expectation and Variance of Imputed Cluster Means

The proof in this section shows that the expected value and variance of the imputed cluster mean ($\tilde{v}_k$) are unbiased estimates of population mean and between-cluster variance.

$$\begin{aligned}
&E(\tilde{v}_k)\\
&= E[\lambda\bar{x}_k + (1-\lambda)\mu + g_k]\\
&= \lambda E(\bar{x}_k) + (1-\lambda)\mu + E(g_k)\\
&= \lambda\mu + (1-\lambda)\mu\\
&= \mu
\end{aligned} \tag{4.1}$$

and

$$\begin{aligned}
&Var(\tilde{v}_k)\\
&= Var[\lambda\bar{x}_k + (1-\lambda)\mu + g_k]\\
&= \lambda^2 Var(\bar{x}_k) + Var(g_k)\\
&= \lambda^2 Var(v_k + \bar{\gamma}_k + \bar{E}_k) + Var(g_k)\\
&= \lambda^2 Var(v_k) + \lambda^2 Var(\bar{\gamma}_k) + \lambda^2 Var(\bar{E}_k) + Var(g_k)\\
&= \lambda^2\sigma_b^2 + \lambda^2\sigma_w^2/I + \lambda^2\sigma_e^2/I + (1-\lambda)\sigma_b^2\\
&= \lambda^2(\sigma_b^2 + (\sigma_w^2 + \sigma_e^2)/I) + (1-\lambda)\sigma_b^2\\
&= \lambda\left(\frac{\sigma_b^2}{(\sigma_b^2 + (\sigma_w^2 + \sigma_e^2)/I)}\right)\left(\sigma_b^2 + (\sigma_w^2 + \sigma_e^2)/I\right) + (1-\lambda)\sigma_b^2\\
&= \lambda\sigma_b^2 + (1-\lambda)\sigma_b^2\\
&= \sigma_b^2
\end{aligned} \tag{4.2}$$

The derivation of the variance of the imputed cluster mean ($\tilde{v}_k$) also shows how the variance components are reflected in this statistic. The cluster level reliability coefficient (λ) represents the shrinkage of the cluster mean estimates based on the posterior estimates, $\lambda\bar{x}_k + (1-\lambda)\mu$, toward the population mean and the level of shrinkage of the variance. By construction, the variance of the random component g_k is the posterior variance, which is equal to the portion of the variance shrunk. More shrinkage toward the population mean corresponds to a relatively larger proportion of between-cluster variance from the random added term. According to the definition of λ ($\lambda \equiv \frac{\sigma_b^2}{\sigma_b^2+(\sigma_w^2+\sigma_e^2)/I}$), relatively larger variance within clusters and measurement error variance correspond to a larger proportion of between-cluster variance from the random component g_k, hence a larger sampling variance of the cluster mean based on Rubin's MI estimates.

The within-cluster sample size *I* is another factor in this formula – a larger cluster size makes λ larger, hence the proportion of variance from the random component g_k smaller.

4.2 Derivation of Variance of Within-Cluster Imputations

Within cluster population can be treated as a population without clustering. The within-cluster variance calculated as follows proves that the within-cluster variance of imputations is an unbiased estimate of the population within-cluster variance σ_w^2.

$$\begin{aligned}
&Var(\tilde{\theta}_{ik} \mid k = k') \\
&= Var[\rho x_{ik'} + (1-\rho)v_{k'} + f_{ik'}] \\
&= \rho^2 Var(x_{ik'}) + 0 + Var(f_{ik'}) \\
&= \rho^2(\sigma_w^2 + \sigma_e^2) + 0 + (1-\rho)\sigma_w^2 \\
&= \rho\sigma_w^2 + (1-\rho)\sigma_w^2 \\
&= \sigma_w^2
\end{aligned} \tag{4.3}$$

The derivation of the variance of the imputed individual scores within a cluster $\left(\tilde{\theta}_{ik} \mid k = k'\right)$ also shows how the variance components are reflected in this statistics.

The within-cluster reliability coefficient (ρ) represents the shrinkage of the estimate of individual scores based on the posterior estimates, $\rho x_{ik} + (1-\rho)\nu_k$, toward the population cluster mean and the level of shrinkage of the variance. By construction, the variance of the random component f_{ik} is the posterior variance, which is equal to the portion of the variance shrunk. More shrinkage toward the population cluster mean corresponds to greater proportion of within-cluster variance from the random added term. According to the definition of ρ ($\rho \equiv \frac{\sigma_w^2}{\sigma_w^2+\sigma_e^2}$), relatively larger measurement error variance correspond to larger proportion of within-cluster variance from the random component f_{ik} , hence a larger sampling variance for the individual scores within clusters based on Rubin's MI estimates.

4.3 Derivation of Mean and Variance of the Imputations for the Clustered Population

This proof shows that the expected value and variance of the imputations are unbiased estimates of the mean and variance of the population with a clustered structure.

$$\begin{aligned}
&E(\tilde{\theta}_{ik}) \\
&= E\{\rho x_{ik} + (1-\rho)[\lambda \bar{x}_k + (1-\lambda)\mu + g_k] + f_{ik}\} \\
&= \rho E(x_{ik}) + (1-\rho)[\lambda E(\bar{x}_k) + (1-\lambda)\mu + E(g_k)] + E(f_{ik}) \qquad (4.4) \\
&= \rho\mu + (1-\rho)\mu + 0 \\
&= \mu
\end{aligned}$$

and

$$
\begin{aligned}
& Var(\tilde{\theta}_{ik}) \\
& = Var\{\rho x_{ik} + (1-\rho)[\lambda \bar{x}_k + (1-\lambda)\mu + g_k] + f_{ik}\} \\
& = \rho^2 Var(x_{ik}) + (1-\rho)^2 Var[\lambda \bar{x}_k + (1-\lambda)\mu + g_k] + Var(f_{ik}) + \\
& \qquad 2\rho(1-\rho)\,\mathrm{cov}[x_{ik}, \lambda \bar{x}_k + (1-\lambda)\mu + g_k] \\
& = \rho^2 Var(x_{ik}) + (1-\rho)^2 Var[\lambda \bar{x}_k + (1-\lambda)\mu + g_k] + Var(f_{ik}) + \\
& \qquad 2\rho(1-\rho)\lambda[\mathrm{var}(x_{ik}) + (I-1)\,\mathrm{cov}(x_{pk}, x_{qk} \mid p \neq q)]/I \\
& = \rho^2(\sigma_b^2 + \sigma_w^2 + \sigma_e^2) + (1-\rho)^2\sigma_b^2 + (1-\rho)\sigma_w^2 + \\
& \qquad 2\rho(1-\rho)\lambda[\sigma_b^2 + \sigma_w^2 + \sigma_e^2 + (n-1)\sigma_b^2]/I \\
& = \rho^2\sigma_b^2 + \rho\sigma_w^2 + (1-\rho)^2\sigma_b^2 + (1-\rho)\sigma_w^2 + 2\rho(1-\rho)\sigma_b^2 \\
& = \sigma_b^2 + \sigma_w^2
\end{aligned}
\tag{4.5}
$$

As shown above, the derivation of the variance of the imputed individual scores in a clustered population combines the results from sections 4.1 and 4.2. The reliability coefficients from each stage (ρ and λ) represent the shrinkage at the stage. More shrinkage corresponds to a higher proportion of variance from the random added term at the corresponding stage. In the next chapters, we will further study the impact of the variance components to the sampling variance of the mean of the imputation in a more complex case, imputation with unknown population mean and cluster means. In this more complex case the analytic results were too unwieldy to derive directly, so a simulation study was conducted.

Chapter 5 : Imputation with Unknown Population Mean and Cluster Means

To achieve the goal of this study, a simulation study was designed and carried out, which not only allowed perfect control of factors under consideration in the estimation procedure, but also made it possible to compare the estimates to the true population values. This chapter describes the methodological framework and the application process of the simulation study in five sections. The first section states the three research questions explicitly. The second section extends the construction of multiple imputations to the case in which neither μ nor v_k s are known. This amounts to adding stages of Bayesian estimation for these higher-level parameters, and drawing random values from the posterior distribution, in the construction of each MI data set. In section 5.3 the generation of multiple data sets of imputed θs under a cluster sample design with equal cluster size using MI is described in detail, including discussions of the manipulated factors, the fixed population and the actual generation of simulated data. Section 5.4 specifies an analysis method based on the simulated data. Finally, Section 5 presents analysis results to address the three research questions. The data generation and analyses were carried out using the *R* language and Microsoft Excel.

5.1 Research Questions of the Simulation Study

The purpose of the simulation study is to examine the properties of estimates of population characteristics obtained from MI in the case of unknown population mean and cluster means. The statistics of interest include the point estimates of the population mean, cluster means, overall population variance, and between-cluster and pooled-within-cluster variances based on the plausible values for true scores. The study addresses the following research questions:

1. How are different amounts of variance reconstruction terms incorporated to construct each set of plausible values to recreate the population properties of the true score?
2. How do the variance components and sample sizes impact the sampling variance of the MI-based estimate for the population mean?
3. What are the relationships between the sampling variance of the estimate of the population mean based on the imputations and those based on observations of the true score and the observed score?

5.2 Construction of Imputations for the Case of Unknown Means

In chapter 4, the population parameters μ, σ_b^2, σ_w^2 and σ_e^2 were assumed to be known for the imputation generated in the random-effects model with measurement errors. These simplifying assumptions allowed us to derive closed forms of the estimates that illustrated the properties of the imputations. The relationships illustrated in these calculations add insight to the structure and meaning of the elements used in the construction of imputations. However, these population-level parameters are always unknown in practice, although they may be estimated from the information in previous research and current data. To investigate the model with unknown population parameters, a Bayesian method was applied in a simulation study. In this part of the research, we took into account that the location parameters, namely the population mean μ and the cluster means ν_k, are not known, while still keeping σ_b^2, σ_w^2 and σ_e^2 known. (In practical applications, these variance components will need to be estimated concurrently or from previous research. Analyses with unknown variance components remain a topic for future study.) Simple closed-form derivations are no longer available, but by using well-chosen simulations we can further demonstrate additional properties of the imputations, and add additional insight for potential users for the CTT case as well as for cases that are more complex and less transparent, such as item response theory models.

In investigating the model with unknown population location parameters, a Bayesian procedure with non-informative prior distribution on these parameters was considered. As Gelman et al. (2004) indicated, by using noninformative prior distributions, inferences are not affected by information external to the current data. This method can be approximated by estimating population parameters based on the observed data.

Specifically, the population mean μ was estimated with a sample mean from the simulated data and the variance of the sample mean was calculated, where the estimate of μ is denoted by $\hat{\mu}$ and the variance of the estimate by $\hat{V}_{\hat{\mu}(x)}$. Then, plausible values for μ were constructed by drawing a value from this posterior distribution for the sample mean in a normal approximation for each pseudo dataset of plausible values. That is, for each pseudo dataset *m*, a random number $\tilde{\mu}_{(m)}$ from $N\left(\hat{\mu}, \hat{V}_{\hat{\mu}(x)}\right)$ was drawn. By doing this, the MI procedure built the appropriate amount of uncertainty in estimating the unknown μ into the construction of the plausible values.

5.3 Data Generation

The simulation study created imputed data sets for a variety of contrasting conditions, with notably different sizes of variance components and sample sizes at different levels of the design. In particular, the following variables were created sequentially by randomly drawing from corresponding distributions: to generate a data set of sampled *X*, we created cluster means of true scores ν_k, individual true scores θ_{ik}, and individual observed scores X_{ik}; to produce sets of imputations based on each data set of *X*, for each of *m* pseudo data sets of plausible values, we created imputed cluster means of individual scores $\tilde{\nu}_{k(m)}$, and imputed individual scores $\tilde{\theta}_{ik(m)}$, where the subscripts indicate the *m*th imputed score for the *i*th simulee in cluster *k*. The entire process described above was repeated a large number of times to create repeated samples by using different random seeds in the random draw at each step of selection. As a result, for each repeated sample, a set of ν_k, θ_{ik}, X_{ik}, $\tilde{\nu}_{k(m)}$ and $\tilde{\theta}_{ik(m)}$ were created and *m* data sets of plausible values were saved in *m* data sets.

5.3.1 Manipulated Factors

When generating simulation data to reflect contrasting conditions in the study, the manipulated factors include four groups: variance components, sample sizes, the number of imputations, and the number of repeated samples. The values of these factors are summarized in the table below. Combinations of variance components and sample sizes are used for each of the three research questions. The number of imputations and number of repeated samples are selected among the conditions to effectively address each research question. Note that the full cross-classification of the factors mentioned above is not used for each research question. The chosen conditions are discussed in more detail in the corresponding sections.

Table 5.1 Manipulated Factors

Factors	**# of Conditions**	**Description**
Variance components	15	$(\sigma_b^2, \sigma_w^2, \sigma_e^2) =$ (1, 1, 1),(100, 1, 1), (1, 100, 1), (1, 1, 100), (100, 100, 1), (100, 1, 100), (1, 100, 100), (100, 100, 100), (4, 1, 1), (1, 4, 1), (1, 1, 4), (4, 4, 1), (4, 1, 4), (1, 4, 4), (4, 4, 4)
Sample sizes		
Number of clusters (*K*)	4	5, 30, 100, 300
Cluster size (*I*)	4	5, 30, 100, 300
# of imputations	2	10, 100
# of repeated samples	3	1000, 5000, 25000

5.3.1.1 Population Variance Components

A wide range of ratios between variance components was selected for the simulation study. This range more than covers commonly observed ratios in social research, by addressing a wide numeric range of the ratio. The ratios are reflected by the values of the components, as the baseline condition sets all three variance components to be equal to one, which is represented by $(\sigma_b^2, \sigma_w^2, \sigma_e^2)$ = (1, 1, 1). Other conditions show inflation of certain component(s), which are represented by $(\sigma_b^2, \sigma_w^2, \sigma_e^2)$ = (100, 1, 1), (1, 100, 1), (1, 1, 100), (100, 100, 1), (100, 1, 100) and (1, 100, 100). The implications of these very different structures for the elements of imputation will be pointed out.

In addition to the extreme values of the ratios shown above, to represent situations commonly seen in practice, the simulation also used a set of moderate values, where the ratio of the variance components is 4. Specifically, the three variance components are set to $(\sigma_b^2, \sigma_w^2, \sigma_e^2)$ = (4, 1, 1), (1, 4, 1), (1, 1, 4), (4, 4, 1), (4, 1, 4) and (1, 4, 4).

All the combinations of variance components were used in the investigation of all three research questions. In examining research question two, this wide range of variance ratios was used to fully evaluate the effect of the relative size of the variance components in the population on Rubin's MI-based estimate of the variance of the population mean estimates. For other research questions, these ratios represent a sufficient coverage of possible situations.

5.3.1.2 Sample Sizes

Following the same scheme as for the variance components, sample sizes in the simulation study were selected to represent a large range of values covering more than normally observed in social research. For example, smaller sample sizes could appear in practice, especially when analyzing sub-domains of the population. To reflect such cases, the minimum sample size is set to 5 at both sampling stages. As this study assumes normal distribution at both levels of the clustered population, sampling distribution with small sample sizes is also normal.

This study used the combinations of sample sizes with 5, 30, 100 and 300 at each sampling stage; that is, K = 5 and I = 5; K = 5 and I = 30; …; K = 30 and I = 5;…; K = 300 and I = 300, where K is the number of clusters and I is the number of observations within each cluster.

These combinations of sample sizes were used for all three research questions. These settings impact the reliability of cluster means and the precision of sample estimates.

5.3.1.3 Number of Imputations

To determine the number of imputations needed for applications, Rubin (1987, p.114) illustrated the relationship between the number of imputations and relative efficiency (RE) of the estimator from MI as follows:

$$RE \equiv \sqrt{\frac{V\left(s_{(\infty)}\left(\widetilde{\theta}_{ik(\infty)} \middle| x_{ik} \right) \right)}{V\left(s_{(m)}\left(\widetilde{\theta}_{ik(m)} \middle| x_{ik} \right) \right)}} = \left(1 + \frac{\gamma_0}{m}\right)^{-\frac{1}{2}} \qquad (5.1)$$

Defined as the efficiency when using a finite number of proper imputations, *m*, rather than an infinite number, *RE* can be expressed as a function of the expected fraction of missing information (γ_0), and the number of imputations (m). $V\left(s_{(\infty)}\left(\widetilde{\theta}_{ik(\infty)} \middle| x_{ik} \right) \right)$ represents the conditional variance of point estimates based on an infinite number of imputations of θ_{ik} given the observed x_{ik} and $V\left(s_{(m)}\left(\widetilde{\theta}_{ik(m)} \middle| x_{ik} \right) \right)$ represents the conditional variance based on *m* imputations.

For point estimates in a large sample, the *RE*s achieved for various values of *m* and rates of missing information are shown in Table 5.2 (Rubin, 1987, p.114).

Table 5.2 Large-sample relative efficiency (in %) in units of standard deviations when using a finite number of proper imputations, *m*, rather than an infinite number, as a function of the fraction of missing information, γ_0.

	γ_0					
m	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**1.0**
3	0.98	0.95	0.93	0.90	0.88	0.87
5	0.99	0.97	0.95	0.94	0.92	0.91
10	1.00	0.99	0.98	0.97	0.96	0.95

In the CTT case, where measurement error is the issue, the concept of the fraction of missing information (γ_0) could be generalized and calculated as the complement of the reliability of the true score versus observed score. The proof is provided in the ETS research memorandum by Robert Mislevy (in press) . In the case of a clustered population, $\rho = \frac{\sigma_b^2+\sigma_w^2}{\sigma_b^2+\sigma_w^2+\sigma_e^2}$. In the most extreme case in this study, where $(\sigma_b^2, \sigma_w^2, \sigma_e^2)$ = (1, 1, 100), the proportion of missing information is close to 1. To gain a RE value over 0.95 for all conditions in the imputation, 10 imputations are needed. This study used 10 imputations for all the research questions, except the case in the next paragraph.

Although Table 5.2 shows that 10 imputations are sufficient for point estimates, more imputations may be needed to estimate the sampling variance of these point estimates for the imputations. To examine the impact of the number of imputations to the estimation of sampling variance of means, for research question three, 100 imputation data sets were created for a selective set of simulations when ratios of variance components were 100 and the number of repeated samples was 5K. The variation of the sampling variance was compared to simulations with 10 imputation data sets.

5.3.1.4 Number of Repeated Samples

To gain an appropriate precision level in the study or compute empirical variance estimates of sample statistics, statistics were computed from samples repeatedly selected from the populations of θ, X and the imputed scores. In the interest of limiting the program running time, the number of repetitions went beyond 1000 only when necessary. For research question one, 1000 repetitions were run, while 5000 were run for research question two. For research question three, different numbers of repeated samples including 1000, 5000, to 25000 were created to examine the convergence of the statistics of interest to the expected patterns in terms of number of repetitions.

5.3.2 Fixed Factor - Population

This simulation study generated the sample data from an infinite population, which approximated the population of interest in this study - a finite population with large population size. The concept of "Superpopulation" can be used to represent this hypothetical infinite population from which the finite population is a sample (Deming & Stephan, 1941). The validation of this approximation to the finite population is discussed by Skinner, Holt, and Smith (1989), who wrote “super-population parameters may often be preferred to finite population parameters as targets of inference in analytic surveys. However, if *n* is large, there will often be little numerical difference between the two.” As the sample was selected in two stages, the large population size is assumed for both the number of clusters and the number of observations within each cluster.

5.3.3 Data Generation

For each combination of all the factors discussed above, the simulation study took two steps in creating the imputation data sets to be used for analysis. First, the observed test score data, *X*s, were created according to the measurement model and the population model as discussed in sections 2.1.1 and 2.3. Then multiple data sets of imputed true score θs were constructed based on the simulated observed data using MI as discussed in section 2.6.4. Step one was carried out with a targeted number of repeated samples. For each such repetition, step two was repeated multiple times to create multiple sets of imputed data sets for a given sample of *X*s, using different random seeds in creating random components.

5.3.3.1 Create Simulated Data on Observed Test Score *X*

To reflect the clustering feature of the sample and the measurement error of the observed score in the classical test theory, the simulated test score *X* was generated in three steps. In the first step, cluster means of the *k^th^* cluster, ν_k, where *k* = 1 to *K*, were randomly drawn from the normal distribution $N\left(0,\sigma_b^2\right)$. Then, the true score θ_{ik} of *ith* student in cluster *k*, where *i* = 1 to *I*, was constructed by adding a randomly selected value from the normal distribution $N\left(0,\sigma_w^2\right)$ to ν_k. Finally, the observed score X_{ik} was formed by adding a random number drawn from the normal distribution $N\left(0,\sigma_e^2\right)$ to θ_{ik}. As shown in Table 5.1, the sample sizes at the two sampling stages, *K* and *I*, both took the same set of values 5, 30, 100 and 300.

The variance components of the population, σ_b^2, σ_w^2 and σ_e^2, represent between-cluster variance, within-cluster variance, and error variance, respectively. To examine a variety of relative size of variance components in the analysis, a baseline simulation data set was created by setting the combination of σ_b^2, σ_w^2 and σ_e^2 to (1, 1, 1). Then, data sets were generated based on the other combinations under consideration: $(\sigma_b^2,\sigma_w^2,\sigma_e^2)$ = (100, 1, 1), (1, 100, 1), (1, 1, 100), (100, 100, 1) , (100, 1, 100), (1, 100, 100), (4, 1, 1), (1, 4, 1), (1, 1, 4), (4, 4, 1) , (4, 1, 4), and (1, 4, 4), where 100 and 4 represent ratios of certain variance components.

5.3.3.2 Generate Imputations Based on Observed Values

We then generated *m* sets of imputations of true scores, $\tilde{\theta}_{ik(m)}$, also known as plausible values, for each simulated data set of observed test score X_{ik} corresponding to the combinations of variance components, where *m* was set to 10 for most parts of the study. This imputation was constructed using a two-level extension of Kelly's formula, as shown in Chapter 3, $\rho x_{ik} + (1-\rho)[\lambda \bar{x}_k + (1-\lambda)\tilde{\mu}_{(m)} + g_{k(m)}] + f_{ik(m)}$, where $\tilde{\mu}_{(m)}$, $g_{k(m)}$ and $f_{ki(m)}$ are random variables across the *m* data sets. Specifically, $\tilde{\mu}_{(m)}$ was randomly drawn from the distribution of the sample mean of the observed *x*'s $N(\hat{\mu}, \hat{V}_{\hat{\mu}(x)})$, $g_{k(m)}$ was drawn from $N(0,(1-\lambda)\sigma_b^2)$, and $f_{ki(m)}$ was drawn from $N(0,(1-\rho)\sigma_w^2)$. The sample cluster mean of cluster *k* is denoted as $\bar{x}_k$. As described in Chapter 3, $\rho = \sigma_w^2 / (\sigma_w^2 + \sigma_e^2)$ and $\lambda = \sigma_b^2 / [\sigma_b^2 + (\sigma_w^2 + \sigma_e^2)/I]$.

The value of *m* was extended to 100 when the ratio of variance components was 100 and the number of repeated samples was 5,000, to examine the impact of a larger number of imputations on the estimation of sampling variance.

5.3.3.3 Repeated Samples

By repeating the steps in sections 5.4.3.1 and 5.4.3.2, samples were repeatedly drawn from the same population distribution in creating repeated samples of the true score θ_{ik} , the observed score X_{ik} and the multiple imputed score $\tilde{\theta}_{ik(m)}$. Different seeds were used in the creation of the random components of each repeated sample.

5.3.3.4 Random Seeds

To generate independent random variables in the simulation study, a different seed value was used for each random component in each repeated sample in the R code. All random variables were selected from normal distributions with the means and variances specified above.

5.4 Analysis

5.4.1 Study Method

The simulation study intends to investigate the three research questions in the following manner.

First, the study demonstrated and discussed how the variance reconstruction terms were reflected in the construction of imputations to recreate the population properties of the true score. Based on the simulation factors in each condition, as described in section 5.3, reliability coefficients, ρ and λ, were computed to show the contribution of the observed score to the imputed score. Then variances of random terms g_k and f_{ik}, also called variance reconstruction terms, were presented to show the contribution to the variance of the imputed scores. By adding these terms to the creation of imputed scores, it was empirically demonstrated that population characteristics were recovered from the imputed score for these substantially different population structures of the observed scores. To show the unbiasedness of sample statistics based on the imputed score $\tilde{\theta}_{ik(m)}$, the empirical distribution of the sample statistics for each simulation condition was constructed by randomly generating the imputed data repeatedly 1000 times, that is, producing 1000 observed data sets ***X***. Let *S* generically denote a population characteristic of θ_{ik}. We empirically calculated the mean of the sample estimates of *S* across the 10 plausible values generated for each ***X***, s_M, and the corresponding sampling variance across the 1000 repeated samples as

$$\bar{s}_M = \sum s_M / 1000 \tag{5.2}$$

and

$$Var(s_M) = \frac{\sum (s_M - \bar{s}_M)^2}{1000 - 1} \tag{5.3}$$

respectively. *Z*-values can be calculated as the standardized score for $\bar{s}_M$, as shown in the formula below:

$$z - value = \frac{\bar{s}_M - S}{\sqrt{Var(s_M)/1000}} \tag{5.4}$$

This *z*-value represents the distance in standard errors between the sample estimates and the population value. To demonstrate the unbiased characteristics of the sample estimates, we calculated *z*-values for the following statistics: the overall sample mean, the cluster sample mean, the overall sample variance, the between-cluster variance, and the pooled within-cluster variance.

Similarly, the same set of *z*-values were computed based on the true score *θ*'s and the observed *x*'s in the sample. The estimate based on the true score *θ* was treated as the gold standard, which is the best estimate one can get from a sample if the individual true score can be observed. The estimates based on the observed score *X* is the Maximum Likelihood estimate, which is an unbiased and efficient estimator for an individual person's true score, but not necessarily for the population distribution. The *z*-values based on imputed scores and the observed *X*'s were evaluated by comparing them to the estimates based on the true score *θ*'s.

Second, this study demonstrates empirically the impact of variance components and sample sizes on the sampling variance of the mean estimate based on Rubin's formula for MI, which includes within-imputation variance U_M, between imputation variance B_M and total variance V_M. Because clear relationships among these estimates can only be detected at a certain precision level of the estimation process, which may not be met with the sample size settings in the simulation conditions, the statistics were estimated for repeated samples to increase the precision of the sample estimates. For each sample, sampling variances of the sample mean were estimated based on theoretical formulas, and then averaged over 5,000 repeated samples. Patterns of the relationship were first explored by comparing the estimates of V_M , U_M , and B_M across the simulation conditions. Graphs were generated for the purpose of illustration. Then, the relationship was further investigated using regression models on V_M , U_M , B_M and the ratio of U_M over V_M , where the simulation factors were treated as independent variables. Main effects, interaction terms and terms in higher order were studied.

Finally, this study demonstrated empirically the property of the sampling variance of the mean estimate based on imputed scores in terms of the relationship to the sampling variance of the estimates based on the true score θ and the observed score *X*. The three sampling variances were compared to each other according to the simulation conditions. As in studying research question two, to show the relationships of these estimates clearly, sampling variances were averaged across the estimates of the 5,000 repeated samples for each simulation condition. To present the convergence to the expected relationship among sampling variances of statistics in terms of the number of repeated samples, the same analysis was also done with different numbers of repeated samples, 1,000 and 25,000, and then the result was compared to that of the 5,000 repeated samples. In addition, we increased the number of imputations from 10 to 100 for the case with 5,000 repeated samples to show the impact on the estimated sampling variance based on imputed scores. In addition, we applied regression models to examine the relationships between the simulation factors and the ratio of the sampling variance of the true score θ to that of the imputed score, which were used as outcome variables in the model. The simulation factors were treated as independent variables. Main effects, interaction terms and other high order terms were studied.

5.4.2 Estimation

As discussed in section 2.6, inferences are made about population characteristics using the multiple imputed data sets from the simulation following Rubin's formulation. Inferences based on *m* imputation datasets start with the calculation of the sample estimates for each dataset, approximated by $N(s_{(m)}, U_{(m)})$, where $s_{(m)}$ is the point estimate of the statistic of interest calculated based on imputation data set *m* and $U_{(m)}$ is the sampling variance of the point estimate treating the imputed values as observed. Then $s_{(m)}$ and $U_{(m)}$ are averaged across the *M* estimates to obtain s_M and U_M. Across the three research questions in this study, the statistics of interest include the population mean and variance for a two-stage cluster sample, where the putatively unbiased estimators are the mean and variance of the imputed data in the sample, denoted as s_M.

Research question one studies a set of sample statistics s_M, which are expected to be unbiased estimators of population statistics *S*. For comparison purposes, parallel estimators based on the true score θ_{ik} and the observed score x_{ik} were also examined. Table 5.3 presents the formulas in calculating these $s_{(m)}$ and the corresponding estimators based on the true score θ_{ik} and the observed score x_{ik}.

The within-cluster variance was estimated with the pooled within-cluster variance based on the imputed score, the true score and the observed score of the sample. The unbiased estimator of the sampling variance of the cluster mean based on the true score was developed as below (Cochran, 1977, p278):

$$
\begin{aligned}
s_1^2 - \frac{1}{I} \times s_2^2 &= Var(\overline{\theta}_k) - \frac{1}{I} \times Var(\theta_{ik} \mid k) \\
&= \frac{\sum_k (\overline{\theta}_k - \overline{\overline{\theta}})^2}{K-1} - \frac{1}{I} \times \frac{\sum_k \sum_i (\theta_{ik} - \overline{\theta}_k)^2}{K(I-1)}
\end{aligned}
\tag{5.5}
$$

where s_1^2 is the estimated between-cluster element variance and s_2^2 is the estimated within-cluster element variance. The estimator based on the imputed score and the observed score were developed in the same way. Finally, the estimator for the total variance was the summation of the two estimators above.

The bias of the point estimates was calculated to show the unbiased character of the sample estimates, where the bias is the difference between the sample estimates and the true population value, $Bias = s_M - S$.

Table 5.3 Population statistics and corresponding estimators based on the imputed scores, the true score and the observed scores

Population statistics (*S*)	**Sample estimator ($s_{(m)}$) based on the imputed score ($\tilde{\theta}_{ik(m)}$)**	**Sample estimator based on the true score (θ_{ik})**	**Sample estimator based on the observed score (X_{ik})**
Overall mean (μ)	$\bar{\bar{\tilde{\theta}}}_{(m)} = \frac{\sum_k \sum_i \tilde{\theta}_{ik(m)}}{K \times I}$	$\bar{\bar{\theta}} = \frac{\sum_k \sum_i \theta_{ik}}{K \times I}$	$\bar{\bar{x}} = \frac{\sum_k \sum_i x_{ik}}{K \times I}$
Cluster means (ν_k)	$\bar{\tilde{\theta}}_{k(m)} = \frac{\sum_i \tilde{\theta}_{ik(m)}}{I}$	$\bar{\theta}_k = \frac{\sum_i \theta_{ik}}{I}$	$\bar{x}_k = \frac{\sum_i x_{ik}}{I}$
Within-cluster variance (σ_w^2)	$Var(\tilde{\theta}_{ik(m)} \mid k) = \frac{\sum_k \sum_i (\tilde{\theta}_{ik(m)} - \bar{\tilde{\theta}}_{k(m)})^2}{K(I-1)}$	$Var(\theta_{ik} \mid k) = \frac{\sum_k \sum_i (\theta_{ik} - \bar{\theta}_k)^2}{K(I-1)}$	$Var(x_{ik} \mid k) = \frac{\sum_k \sum_i (x_{ik} - \bar{x}_k)^2}{K(I-1)}$
Variance of cluster means (σ_b^2)	$Var(\bar{\tilde{\theta}}_{k(m)}) - \frac{1}{I} \times Var(\tilde{\theta}_{ik(m)} \mid k)$ $= \frac{\sum_k (\bar{\tilde{\theta}}_{k(m)} - \bar{\bar{\tilde{\theta}}}_{(m)})^2}{K-1} - \frac{1}{I} \times \frac{\sum_k \sum_i (\tilde{\theta}_{ik(m)} - \bar{\tilde{\theta}}_{k(m)})^2}{K \times (I-1)}$	$Var(\bar{\theta}_k) - \frac{1}{I} \times Var(\theta_{ik} \mid k)$ $= \frac{\sum_k (\bar{\theta}_k - \bar{\bar{\theta}})^2}{K-1} - \frac{1}{I} \times \frac{\sum_k \sum_i (\theta_{ik} - \bar{\theta}_k)^2}{K(I-1)}$	$Var(\bar{x}_k) - \frac{1}{I} \times Var(x_{ik} \mid k)$ $= \frac{\sum_k (\bar{x}_k - \bar{\bar{x}})^2}{K-1} - \frac{1}{I} \times \frac{\sum_k \sum_i (x_{ik} - \bar{x}_k)^2}{K(I-1)}$
Total variance ($\sigma_b^2 + \sigma_w^2$)	$Var(\tilde{\theta}_{ik}) = Var(\bar{\tilde{\theta}}_{k(m)}) + (1 - \frac{1}{I}) \times Var(\tilde{\theta}_{ik(m)} \mid k)$ $= \frac{\sum_k (\bar{\tilde{\theta}}_{k(m)} - \bar{\bar{\tilde{\theta}}}_{(m)})^2}{K-1} + \frac{\sum_k \sum_i (\tilde{\theta}_{ik(m)} - \bar{\tilde{\theta}}_{k(m)})^2}{K \times I}$	$Var(\theta_{ik}) = Var(\theta_{ik} \mid k) + Var(\bar{\theta}_k)$ $= \frac{\sum_k (\bar{\theta}_k - \bar{\bar{\theta}})^2}{K-1} + \frac{\sum_k \sum_i (\theta_{ik} - \bar{\theta}_k)^2}{K \times I}$	$Var(x_{ik}) = Var(x_{ik} \mid k) + Var(\bar{x}_k)$ $= \frac{\sum_k (\bar{x}_k - \bar{\bar{x}})^2}{K-1} + \frac{\sum_k \sum_i (x_{ik} - \bar{x}_k)^2}{K \times I}$

Note: *K* is the number of clusters in the sample and *I* is the sample size within cluster

For research questions two and three, the sampling variance $U_{(m)}$ of the sample mean of the imputed score was estimated for each imputed data set. Treating the $\tilde{\theta}_{ik(m)}$ values as if they were values of $\theta_{ik(m)}$ observed directly from a corresponding population, the estimator of the sampling variance of the sample mean for imputation *m* is $Var(\overline{\overline{\tilde{\theta}}}_{(m)})$, which can also be denoted as $U_{(m)}$,

$$Var(\overline{\overline{\tilde{\theta}}}_{(m)}) = \frac{1-f_1}{K} s_1^2 + \frac{f_1(1-f_2)}{K \times I} s_2^2,$$

where f_1 and f_2 are finite population correction factors at the two sampling stages.

Using the notation in this study, the formula can be re-written as:

$$Var(\overline{\overline{\tilde{\theta}}}_{(m)}) = \frac{1-\frac{K}{\boldsymbol{K}}}{K} s_1^2 + \frac{\frac{K}{\boldsymbol{K}}(1-\frac{I}{\boldsymbol{I}})}{KI} s_2^2,$$

where *K* represents the number of sampled clusters among the $\boldsymbol{K}$ clusters in the population, *I* is the number of sampled persons among the $\boldsymbol{I}$ persons in the cluster in the population. In the case of this study, since $\boldsymbol{K}$ and $\boldsymbol{I}$ are both infinite numbers, the formula is simplified as

$$U_{(m)} = Var(\overline{\overline{\tilde{\theta}}}_{(m)}) = \frac{1}{K} s_1^2 = \frac{\sum_k (\overline{\tilde{\theta}}_{k(m)} - \overline{\overline{\tilde{\theta}}}_{(m)})^2}{K(K-1)} \tag{5.6}$$

where $\overline{\tilde{\theta}}_{k(m)}$ is the cluster mean for the imputation data set *m* and $\overline{\overline{\tilde{\theta}}}_{(m)}$ is the population mean of that imputation data set.

The statistics computed from the *m* imputed data sets were combined to gain the multiple imputation inference, as shown in formula (2.16). The following statistics can be calculated: the overall estimate of the population statistics of interest

s_M, the average within imputation sampling variance U_M of the estimate, the between imputation variance B_M and the total variance V_M.

The sampling variance for the true score and the observed score can be denoted as

$$Var(\bar{\bar{\theta}}) = \frac{\sum_k (\bar{\theta}_k - \bar{\bar{\theta}})^2}{K(K-1)} \text{ and } Var(\bar{\bar{x}}) = \frac{\sum_k (\bar{x}_k - \bar{\bar{x}})^2}{K(K-1)},$$

where $\bar{\theta}_k$ is the mean of cluster *k* for the true score and $\bar{\bar{\theta}}$ is the overall mean of the true score, while $\bar{x}_k$ is the mean of cluster *k* for the observed score and $\bar{\bar{x}}$ is the overall mean of the observed score. $Var(\bar{\bar{x}})$ is expected to be no less than $Var(\bar{\bar{\theta}})$ since

$$Var(\bar{\bar{x}}) = \frac{\sum_k (\bar{x}_k - \bar{\bar{x}})^2}{K(K-1)} = \frac{\sum_k (\bar{\theta}_k - \bar{\bar{\theta}})^2}{K(K-1)} + \frac{\sum_k (\bar{e}_k - \bar{\bar{e}})^2}{K(K-1)} = Var(\bar{\bar{\theta}}) + Var(\bar{\bar{e}}) \qquad (5.7)$$

and $Var(\bar{\bar{e}})$ is larger than or equal to zero. The relationship between V_M and $Var(\bar{\bar{x}})$ is not clear and will be explored in the simulation study.

In addition, as shown below, according to Cochran (1977, p.278), both $Var(\bar{\bar{\theta}})$ and U_M have the expected value of $\frac{\sigma_b^2}{K} + \frac{\sigma_w^2}{KI}$ when the population size is infinite.

$$E\left(Var(\bar{\bar{\theta}})\right) = E\left(\frac{s_1^2}{K}\right) = \frac{1}{K} \times \left(Var\left(\bar{\theta}_k\right) + \frac{Var\left(\theta_{ik} \mid k\right)}{I}\right) = \frac{\sigma_b^2}{K} + \frac{\sigma_w^2}{KI} \qquad (5.8)$$

$$\begin{aligned} E\left(U_M\right) &= E\left(\frac{\sum U_{(m)}}{M}\right) = E\left(\frac{\sum Var(\bar{\bar{\tilde{\theta}}}_{(m)})}{M}\right) = E\left(\frac{s_1^2}{K}\right) \\ &= \frac{1}{K} \times \left(Var\left(\bar{\tilde{\theta}}_k\right) + \frac{Var\left(\tilde{\theta}_{ik} \mid k\right)}{I}\right) = \frac{\sigma_b^2}{K} + \frac{\sigma_w^2}{KI} \end{aligned} \qquad (5.9)$$

5.5 Results

5.5.1 Research Question 1: How are different amounts of variance reconstruction terms incorporated in the plausible values to recreate the population properties of the true score?

This section starts with illustrating intuitively how the variance reconstruction terms are reflected in the imputation statistics using the simulation data, which is theoretically discussed in Chapter 4. Then, it will be shown that, by incorporating the variance reconstruction terms, each set of plausible values recreates the population means and variances under a two-stage sample design.

The posterior mean of θ_{ik}, $\rho x_{ik} + (1-\rho)[\lambda \bar{x}_k + (1-\lambda)\tilde{\mu}_{(m)}]$, shrinks towards the population mean at the cluster level and shrinks towards the cluster mean at the individual level. The variance of the posterior mean then becomes lower than the variance of the mean of the true score. For each set of plausible values, $\rho x_{ik} + (1-\rho)[\lambda \bar{x}_k + (1-\lambda)\tilde{\mu}_{(m)} + g_{k(m)}] + f_{ik(m)}$, where m = 1, …, 10, the random variance reconstruction terms $g_{k(m)}$ and $f_{ik(m)}$ are included to inflate the variance of the imputed score back to the variance of the true score while keeping the mean estimates unbiased, where $g_{k(m)}$ reflects the posterior variance of the cluster mean estimate and $f_{ik(m)}$ reflects the posterior variance of the individual score estimate. Given that $Var(g_k) = (1-\lambda)\sigma_b^2$ and $Var(f_{ik}) = (1-\rho)\sigma_w^2$, where the reliability coefficients λ and ρ are defined as below,

$$\lambda = \frac{\sigma_b^2}{\left[\sigma_b^2 + \dfrac{(\sigma_w^2 + \sigma_e^2)}{I}\right]}$$

$$\text{and } \rho = \frac{\sigma_w^2}{(\sigma_w^2 + \sigma_e^2)},$$

the posterior variance can be expressed as

$$Var(g_k) = \frac{\sigma_b^2 \times \frac{(\sigma_w^2 + \sigma_e^2)}{I}}{\left[\sigma_b^2 + \frac{(\sigma_w^2 + \sigma_e^2)}{I}\right]} \tag{5.10}$$

$$\text{and } Var(f_{ik}) = \frac{\sigma_w^2 \times \sigma_e^2}{(\sigma_w^2 + \sigma_e^2)} \tag{5.11}$$

For a better understanding of the contribution of the variance reconstruction term, we calculated the relative amount of variance accounted for by the variance reconstruction terms, that is, the proportion of the posterior variance over the variance of the imputed score, denoted as $R\_Var(g_k)$ and $R\_Var(f_{ik})$, and calculated as follows:

$$R\_Var(g_k) = \frac{Var(g_k)}{Var(\tilde{\theta}_{ik})} = \frac{\sigma_b^2 \times \frac{\left(\sigma_w^2 + \sigma_e^2\right)}{I}}{\left[\sigma_b^2 + \frac{\left(\sigma_w^2 + \sigma_e^2\right)}{I}\right] \times \left(\sigma_b^2 + \sigma_w^2\right)} \tag{5.12}$$

$$\text{and } R\_Var(f_{ik}) = \frac{Var(f_{ik})}{Var(\tilde{\theta}_{ik})} = \frac{\sigma_w^2 \times \sigma_e^2}{\left(\sigma_w^2 + \sigma_e^2\right) \times \left(\sigma_b^2 + \sigma_w^2\right)} \tag{5.13}$$

The simulation results are presented for two groups of simulation conditions where the relative sizes of the variance components are 100 and 4, respectively. Table 5.4 and Table 5.5 present the simulation factors: the variance components and sample sizes, reliability coefficients, variances of the reconstruction terms, and the relative variances. These statistics come directly from the settings of the simulation factors and reflect the population characteristics. Since the sample in the simulation study was drawn from a normal distribution with the population variance for these terms, the expected variance of the sample is equal to the population statistics. Note that the number of clusters as a simulation factor is not shown in the table as it does not affect any of the statistics in the table.

As shown in Table 5.4, for the case with extreme ratios of variance components (100), larger values of the posterior variance $Var(g_k)$ correspond to larger values of any of the three variance components, between-cluster variance (σ_b^2), within-cluster variance (σ_w^2) and measurement error variance (σ_e^2). Larger values of σ_w^2 and σ_e^2 correspond to smaller values of λ and larger proportions of σ_b^2 added from the random component g_k to the variance of the imputed score. The relative variance $R\_Var(g_k)$ is larger when σ_e^2 is larger, and when σ_w^2 is smaller, controlling other factors constant, with one exception – a larger $R\_Var(g_k)$ is obtained when σ_w^2 is larger for the cases with σ_b^2 = 100 and σ_e^2 = 1. In addition, a larger $R\_Var(g_k)$ is obtained when σ_b^2 is smaller for the cases with σ_w^2 = 1. On the other hand, a larger $R\_Var(g_k)$ is obtained when σ_b^2 is larger for the cases with σ_w^2 = 100, except when the cluster size is extremely large (300). Among all the combinations of the variance components, the combination (σ_b^2 ,σ_w^2, σ_e^2)= (1,1,100) corresponds to the largest values of $R\_Var(g_k)$ and the combination (σ_b^2 ,σ_w^2, σ_e^2) = (100,1,1) corresponds to the smallest values, given the cluster size.

Holding the values of all the variance components constant, a larger cluster size corresponds to a smaller variance of the random component g_k and a smaller relative variance. When the cluster size is 5 and (σ_b^2 ,σ_w^2, σ_e^2) = (1,1,100), $R\_Var(g_k)$ has the largest value (47.6%) in the table. Among all cases, $Var(g_k)$

accounts for more than 5% of the overall variance only in rare cases, either when (σ_b^2 , σ_w^2, σ_e^2) = (1,1,100) or when cluster size is 5.

In constructing individual scores, a larger variance of f_{ik} is associated with larger values of σ_e^2 and σ_w^2. A larger values of σ_e^2 corresponds to a smaller value of ρ and a larger proportion of σ_w^2 added from the random component f_{ik} to the variance of the imputed score. A larger relative variance $R\_Var\left(f_{ik}\right)$ is associated with a larger value of σ_e^2 and a smaller value of σ_b^2. When σ_b^2 = 1 and σ_e^2 = 1, a larger relative variance is obtained when σ_w^2 decreases from 100 to 1. When σ_b^2 = 100 and σ_e^2 = 100, the opposite relationship is observed – a larger relative variance is obtained when σ_w^2 increases from 1 to 100. For other combinations of σ_b^2 and σ_e^2, σ_w^2 has little impact on relative variance.

$Var\left(f_{ik}\right)$ accounts for a large percentage of the overall variance for the following cases: 25% when $\sigma_b^2 = \sigma_w^2 = \sigma_e^2$ and 49.5% when σ_b^2 = 1 and σ_e^2 = 100. The smallest percentage is 0.5% when σ_b^2 = 100 and σ_e^2 = 1.

Table 5.4 The variance of the variance reconstruction terms by simulation factors when the ratio of the variance components is 100.

Simulation Factors				**Derived Statistics**		**Variance of Variance Reconstruction Terms**				
σ_b^2	σ_w^2	σ_e^2	**Cluster Size (*I*)**	***λ***	***ρ***	***Var(g$_k$)***	***Var(f$_{ik}$)***	**Expected Var($\tilde{\theta}_{ik}$)**	***R_Var(g$_k$)***	***R_Var (f$_{ik}$)***
1	1	1	5	0.71	0.50	0.29	0.50	2	0.143	0.250
1	1	1	30	0.94	0.50	0.06	0.50	2	0.031	0.250
1	1	1	100	0.98	0.50	0.02	0.50	2	0.010	0.250
1	1	1	300	0.99	0.50	0.01	0.50	2	0.003	0.250
1	1	100	5	0.05	0.01	0.95	0.99	2	0.476	0.495
1	1	100	30	0.23	0.01	0.77	0.99	2	0.385	0.495
1	1	100	100	0.50	0.01	0.50	0.99	2	0.251	0.495
1	1	100	300	0.75	0.01	0.25	0.99	2	0.126	0.495
1	100	1	5	0.05	0.99	0.95	0.99	101	0.009	0.010
1	100	1	30	0.23	0.99	0.77	0.99	101	0.008	0.010
1	100	1	100	0.50	0.99	0.50	0.99	101	0.00498	0.010
1	100	1	300	0.75	0.99	0.25	0.99	101	0.00249	0.010
1	100	100	5	0.02	0.50	0.98	50	101	0.010	0.495
1	100	100	30	0.13	0.50	0.87	50	101	0.009	0.495
1	100	100	100	0.33	0.50	0.67	50	101	0.007	0.495
1	100	100	300	0.60	0.50	0.40	50	101	0.004	0.495
100	1	1	5	0.996	0.50	0.40	0.50	101	0.004	0.005
100	1	1	30	0.999	0.50	0.07	0.50	101	0.001	0.005
100	1	1	100	0.9998	0.50	0.02	0.50	101	0.0002	0.005
100	1	1	300	0.9999	0.50	0.01	0.50	101	0.0001	0.005
100	1	100	5	0.83	0.01	16.81	0.99	101	0.166	0.010
100	1	100	30	0.97	0.01	3.26	0.99	101	0.032	0.010
100	1	100	100	0.99	0.01	1.00	0.99	101	0.00990	0.010
100	1	100	300	0.997	0.01	0.34	0.99	101	0.00332	0.010
100	100	1	5	0.83	0.99	16.81	0.99	200	0.084	0.005
100	100	1	30	0.97	0.99	3.26	0.99	200	0.016	0.005
100	100	1	100	0.99	0.99	1.00	0.99	200	0.00500	0.005
100	100	1	300	0.997	0.99	0.34	0.99	200	0.00168	0.005
100	100	100	5	0.71	0.5	28.57	50	200	0.143	0.250
100	100	100	30	0.94	0.5	6.25	50	200	0.031	0.250
100	100	100	100	0.98	0.5	1.96	50	200	0.00980	0.250
100	100	100	300	0.99	0.5	0.66	50	200	0.00331	0.250

Table 5.5 presents the case with moderate ratios of variance components (4). The table illustrates similar patterns to table 5.4 except a few cases – the relative variance $R\_Var(g_k)$ decreases when σ_b^2 increases for more cases, where σ_w^2 = 4 and the cluster size is 30, 100 or 300.

When the cluster size is 5 and ($\sigma_b^2, \sigma_w^2, \sigma_e^2$)= (1,1,4), $R\_Var(g_k)$ has the largest value (25.0%) in the table. $Var(g_k)$ accounts for more than 5% of the overall variance only in rare cases, either when ($\sigma_b^2, \sigma_w^2, \sigma_e^2$) = (1,1,4) and cluster size is 30 or when cluster size is 5.

$Var(f_{ik})$ accounts for a large percentage of the overall variance for the following cases: 25% when $\sigma_b^2 = \sigma_w^2 = \sigma_e^2$ and 40.0% when σ_b^2 = 1 and σ_e^2 = 4. The smallest percentage is 10% when σ_b^2 = 4 and σ_e^2 = 1.

Table 5.5 The variance of the variance reconstruction terms by simulation factors when the ratio of the variance components is 4.

Simulation Factors				**Derived Statistics**		**Vrariance of Variance Reconstruction Terms**				
σ_b^2	σ_w^2	σ_e^2	**Cluster Size (*I*)**	λ	ρ	***Var(g_k)***	***Var(f_ik)***	**Expected Var($\tilde{\theta}_{ik}$)**	***R_Var(g_k)***	***R_Var(f_ik)***
1	1	1	5	0.71	0.5	0.29	0.5	2	0.143	0.25
1	1	1	30	0.94	0.5	0.06	0.5	2	0.031	0.25
1	1	1	100	0.98	0.5	0.02	0.5	2	0.010	0.25
1	1	1	300	0.99	0.5	0.01	0.5	2	0.003	0.25
1	1	4	5	0.50	0.2	0.50	0.8	2	0.250	0.40
1	1	4	30	0.86	0.2	0.14	0.8	2	0.071	0.40
1	1	4	100	0.95	0.2	0.05	0.8	2	0.024	0.40
1	1	4	300	0.98	0.2	0.02	0.8	2	0.008	0.40
1	4	1	5	0.50	0.8	0.50	0.8	5	0.100	0.16
1	4	1	30	0.86	0.8	0.14	0.8	5	0.029	0.16
1	4	1	100	0.95	0.8	0.05	0.8	5	0.010	0.16
1	4	1	300	0.98	0.8	0.02	0.8	5	0.003	0.16
1	4	4	5	0.38	0.5	0.62	2	5	0.123	0.40
1	4	4	30	0.79	0.5	0.21	2	5	0.042	0.40
1	4	4	100	0.93	0.5	0.07	2	5	0.015	0.40
1	4	4	300	0.97	0.5	0.03	2	5	0.005	0.40
4	1	1	5	0.91	0.5	0.36	0.5	5	0.073	0.10
4	1	1	30	0.98	0.5	0.07	0.5	5	0.013	0.10
4	1	1	100	1.00	0.5	0.02	0.5	5	0.0040	0.10
4	1	1	300	1.00	0.5	0.01	0.5	5	0.0013	0.10
4	1	4	5	0.80	0.2	0.80	0.8	5	0.160	0.16
4	1	4	30	0.96	0.2	0.16	0.8	5	0.032	0.16
4	1	4	100	0.99	0.2	0.05	0.8	5	0.010	0.16
4	1	4	300	1.00	0.2	0.02	0.8	5	0.003	0.16
4	4	1	5	0.80	0.8	0.80	0.8	8	0.100	0.10
4	4	1	30	0.96	0.8	0.16	0.8	8	0.020	0.10
4	4	1	100	0.99	0.8	0.05	0.8	8	0.006	0.10
4	4	1	300	1.00	0.8	0.02	0.8	8	0.002	0.10
4	4	4	5	0.71	0.5	1.14	2	8	0.143	0.25
4	4	4	30	0.94	0.5	0.25	2	8	0.031	0.25
4	4	4	100	0.98	0.5	0.08	2	8	0.010	0.25
4	4	4	300	0.99	0.5	0.03	2	8	0.003	0.25

After incorporating the variance reconstruction terms, the point estimates of population means and variances based on the imputations are unbiased. Sampling variances of these point estimates were calculated empirically across the 1,000 repeated samples based on the imputations. *Z*-values of the point estimates were then derived for the means over the 1000 repetitions using formula (5.10) shown in section 5.4.1 and presented in Table 5.6. Considering the combination of the factors σ_b^2 ,σ_w^2, σ_e^2 , the number of clusters *(K*), and the cluster size (*I*), as presented in table 5.1, the calculation was carried out for the 128 combinations and summarized for each population statistic and ratio of variance components. The minimum values, means and maximum values of the *z*-values across the 128 combinations are presented.

In general, Table 5.6 shows similar patterns for the cases with ratios of the variance components equal to 100 and the cases with the ratios equal to 4.

For all the statistics, the *z*-values calculated based on the imputed data are close to zero, specifically, between -3.17 and 2.98. The means of the *z*-values across the 128 combinations range between -0.18 and 0.12 for all population statistics and ratios of variance components.

The *z*-values based on the true score θ_{ik} represent the best sample estimates that one can get in the case that the student's true score can be observed. The range of the *z*-values for θ_{ik} (from -3.31 to 2.90) and the range of the mean of *z*-values across the 128 combinations (from -0.20 to 0.15) are on the same scale as those based on the imputed data, which demonstrates the unbiased character of the estimators based on the imputed data. Even with extreme conditions, the underlying rationale of plausible values with random-effects models still produces the unbiased estimates of population characteristics.

The *z*-values were also calculated based on the maximum likelihood estimates of individual scores, the observed score X_{ik}. As shown in Table 5.6, the estimates of the overall mean and the cluster means are unbiased, while the estimates of the within-cluster variance and the total variance could be severely overestimated. This bias may be ignorable in special cases, when the error variance is much smaller than the variance of the true score. Table 5.7 shows the cases where the *z*-value is between -4 and 4 and the bias may be ignored. Interestingly, the estimates of the variance of cluster means are close to the population statistics for all the factor settings, even when the ratio of the variance components is extremely high.

In summary, the *z*-values show that the sample estimator based on the imputed score is unbiased in estimating the population mean, cluster means, total variance, within-cluster variance and between-cluster variance. In contrast, the sample estimator based on the observed score is positively biased in estimating total variance and within-cluster variance, while the bias may be ignorable in rare cases.

Table 5.6 Range and mean of the *z*-value of the point estimates based on the imputed score, the true score and the observed score, across the 128 combinations of simulation factors

Population statistics (*S*)	Statistics from 1000 repeated samples	Sample estimator (s_M) based on imputed score ($\tilde{\theta}_{ik}$)		Sample estimator based on true score (θ_{ik})		Sample estimator based on observed score (X_{ik})	
Ratio of variance components		100	4	100	4	100	4
Overall mean	Minimum	-2.22	-2.41	-2.00	-2.47	-2.16	-2.46
	Maximum	2.75	2.25	2.52	2.25	2.66	2.16
	Mean	-0.01	0.06	-0.03	0.05	-0.01	0.06
Cluster means	Minimum	-2.16	-2.55	-2.63	-2.96	-2.28	-2.86
	Maximum	2.52	1.95	2.47	2.90	2.29	2.42
	Mean	0.05	-0.10	-0.20	0.09	-0.11	0.08
Total variance	Minimum	-2.09	-2.74	-1.97	-3.17	0.18	7.19
	Maximum	2.13	2.56	2.33	2.55	6320.75	1464.75
	Mean	-0.005	-0.11	0.08	-0.15	382.64	219.25
Variance of cluster means	Minimum	-2.37	-2.62	-2.01	-2.75	-2.01	-2.67
	Maximum	2.11	2.59	2.38	2.55	2.14	2.57
	Mean	-0.06	-0.06	-0.003	-0.10	-0.08	-0.07
Within-cluster variance	Minimum	-2.96	-3.17	-2.27	-3.31	0.24	19.08
	Maximum	2.64	2.98	2.25	2.61	6746.69	5350.80
	Mean	0.12	-0.18	0.15	-0.17	850.37	849.93

Table 5.7 List of *z*-values of the estimates where the bias of the estimates based on the observed score may be negligible (*z*-values between -4 and 4)

σ_b^2	σ_w^2	σ_e^2	# of Clusters (*K*)	Cluster Size (*I*)	Total variance of X_{ik}	Within-cluster variance of X_{ik}
100	100	1	30	300	**0.18**	22.04
100	1	1	30	5	**0.22**	121.55
100	1	1	5	300	**0.23**	422.34
100	1	1	30	300	**0.44**	1089.05
100	100	1	30	5	**0.53**	**1.67**
100	100	1	5	30	**0.60**	4.77
100	100	1	5	100	**0.70**	5.69
100	100	1	5	300	**0.72**	8.40
100	100	1	30	100	**0.73**	10.65
100	1	1	30	100	**0.93**	606.42
100	100	1	5	5	**1.00**	**0.75**
100	1	1	5	30	**1.01**	133.55
1	100	1	5	5	**1.07**	**0.24**
100	1	1	5	5	**1.50**	50.43
100	1	1	100	5	**1.52**	230.72
100	100	1	30	30	**1.67**	7.16
100	1	1	5	100	**1.70**	248.91
100	1	1	100	30	**1.76**	598.81
100	100	1	100	300	**1.79**	36.44
100	100	1	300	5	**1.85**	8.55
100	100	1	100	30	**1.94**	11.92
100	100	1	300	300	**2.30**	66.72
100	1	1	100	300	**2.87**	1939.86
100	100	1	100	5	**3.04**	5.54
1	100	1	30	5	**3.17**	**3.27**
100	100	1	100	100	**3.24**	21.39
100	1	1	30	30	**3.26**	344.53
100	1	1	300	30	**3.43**	1090.70
100	100	1	300	30	**3.49**	20.98
100	1	1	300	100	**3.57**	1889.23
100	1	1	100	100	**3.78**	1100.39
100	1	1	300	300	**3.89**	3312.67
1	100	1	5	100	**3.99**	4.35

5.5.2 Research Question 2: How do the variance components and sample sizes impact the estimation error of the MI-based estimate for the population mean?

When examining the sampling variances, we focus on the estimator of the population mean. Using Rubin's formula as shown in formula (2.16), for each combination of simulation factors, we calculated the estimates of sampling variances of statistics from the multiple imputed data, which include the within imputation variance U_M, the between imputation variance B_M and the total variance V_M, and investigated their relationships to each of the simulation factors, that is, variance components and sample sizes at both sampling stages. The sampling variances were computed for each repeated sample and then averaged across 5,000 repeated samples.

The analysis results are discussed as follows for two simulation settings, where the ratio of the variance components is 100 and 4, respectively.

5.5.2.1 The ratio of the variance components is 100

In the analysis in this section, the variance components (σ_b^2, σ_w^2 and σ_e^2) take values 1 or 100.

Tabulation of sampling variances by simulation factors

Tables 5.8 – 5.10 present the statistics V_M, U_M and B_M, respectively, by the simulation factors, where the row variables are the combination of the number of clusters and the cluster size and the column variables are the combination of the three variance components. According to table 5.8, V_M increases when the sample size decreases, including both the cluster size *I* and the number of clusters *K*, and when either of the variance components increases. Table 5.9 shows similar general patterns for the relationship between U_M and the simulation factors, except that σ_e^2 doesn't seem to have any impact on U_M. Table 5.10 also shows similar general patterns for the relationship between B_M and the simulation factors except that σ_b^2 and σ_w^2 don't seem to have any impact on B_M. Some violations of the general patterns are due to sample variation.

Table 5.8 V_M for the plausible values by the simulation factors.

Sample Size		Variance Components $\sigma_b^2/\sigma_w^2/\sigma_e^2$							
K	*I*	1/1/1	1/1/100	1/100/1	1/100/100	100/1/1	100/1/100	100/100/1	100/100/100
5	5	0.279	4.717	4.251	8.576	19.857	24.629	24.417	28.427
5	30	0.213	0.945	0.863	1.605	20.178	20.724	20.993	21.509
5	100	0.203	0.419	0.402	0.623	19.654	19.981	20.409	20.424
5	300	0.201	0.273	0.269	0.342	19.933	20.476	20.017	20.229
30	5	0.047	0.78	0.71	1.42	3.343	4.088	3.998	4.738
30	30	0.036	0.16	0.15	0.26	3.331	3.462	3.457	3.549
30	100	0.034	0.071	0.067	0.104	3.329	3.377	3.354	3.403
30	300	0.033	0.045	0.045	0.057	3.344	3.346	3.340	3.368
100	5	0.014	0.232	0.212	0.429	1.002	1.220	1.198	1.420
100	30	0.0107	0.047	0.044	0.081	0.998	1.036	1.033	1.066
100	100	0.0102	0.021	0.020	0.031	1.004	1.009	1.009	1.021
100	300	0.0101	0.014	0.013	0.017	0.999	1.001	1.005	1.011
300	5	0.0047	0.077	0.071	0.144	0.335	0.407	0.402	0.473
300	30	0.0036	0.016	0.015	0.027	0.33336	0.3457	0.3450	0.3564
300	100	0.0034	0.0070	0.0067	0.0103	0.33285	0.3371	0.3371	0.3400
300	300	0.0033	0.0046	0.0045	0.0057	0.33341	0.3343	0.3341	0.3362

Table 5.9 U_M for the plausible values by the simulation factors.

Sample Size		Variance Components $\sigma_b^2/\sigma_w^2/\sigma_e^2$							
K	*I*	1/1/1	1/1/100	1/100/1	1/100/100	100/1/1	100/1/100	100/100/1	100/100/100
5	5	0.2361	0.2393	4.2062	4.1794	19.8134	20.1793	24.3729	23.9974
5	30	0.2058	0.2065	0.8555	0.8707	20.1705	19.9956	20.9852	20.7686
5	100	0.2004	0.2019	0.3997	0.4005	19.6521	19.7594	20.4070	20.2056
5	300	0.2006	0.1998	0.2679	0.2681	19.9325	20.4029	20.0165	20.1555
30	5	0.04013	0.04004	0.70186	0.69926	3.33569	3.34817	3.99055	4.00649
30	30	0.03443	0.03445	0.14401	0.14405	3.33020	3.33919	3.45586	3.42571
30	100	0.03371	0.03380	0.06677	0.06682	3.32828	3.34015	3.35325	3.36661
30	300	0.03337	0.03328	0.04457	0.04443	3.34372	3.33369	3.33992	3.35534
100	5	0.01201	0.01200	0.20943	0.20964	0.99932	0.99867	1.19535	1.19884
100	30	0.01033	0.01035	0.04337	0.04341	0.99722	0.99929	1.03283	1.02938
100	100	0.01010	0.01010	0.02006	0.02002	1.00339	0.99829	1.00934	1.00976
100	300	0.01004	0.01003	0.01334	0.01334	0.99872	0.99754	1.00479	1.00715
300	5	0.00400	0.00400	0.07004	0.06996	0.33378	0.33388	0.40092	0.39975
300	30	0.00344	0.00344	0.01441	0.01446	0.33324	0.33351	0.34485	0.34424
300	100	0.00337	0.00337	0.00666	0.00667	0.33281	0.33350	0.33706	0.33632
300	300	0.00334	0.00334	0.00444	0.00445	0.33340	0.33308	0.33410	0.33498

Table 5.10 B_M for the plausible values by the simulation factors.

Sample Size		Variance Components $\sigma_b^2 / \sigma_w^2 / \sigma_e^2$							
K	*I*	1/1/1	1/1/100	1/100/1	1/100/100	100/1/1	100/1/100	100/100/1	100/100/100
5	5	0.03944	4.0706	0.04031	3.99660	0.03980	4.04525	0.03990	4.02707
5	30	0.00657	0.6713	0.00671	0.66739	0.00661	0.66205	0.00666	0.67323
5	100	0.00200	0.1974	0.00199	0.20210	0.00199	0.20178	0.00202	0.19864
5	300	0.00067	0.0667	0.00067	0.06701	0.00067	0.06618	0.00067	0.06680
30	5	0.00664	0.6682	0.00664	0.65891	0.00668	0.67238	0.00660	0.66534
30	30	0.00111	0.1107	0.00111	0.10935	0.00111	0.11202	0.00111	0.11189
30	100	0.00033	0.0337	0.00033	0.03350	0.00033	0.03375	0.00033	0.03285
30	300	0.00011	0.0111	0.00011	0.01116	0.00011	0.01113	0.00011	0.01115
100	5	0.00200	0.1997	0.00200	0.19964	0.00199	0.20084	0.00200	0.20102
100	30	0.00033	0.0335	0.00033	0.03374	0.00033	0.03317	0.00034	0.03301
100	100	0.00010	0.0100	0.00010	0.01002	0.00010	0.00997	0.00010	0.01003
100	300	0.00003	0.0033	0.00003	0.00335	0.00003	0.00333	0.00003	0.00333
300	5	0.00066	0.0664	0.00067	0.06712	0.00068	0.06634	0.00066	0.06670
300	30	0.00011	0.0112	0.00011	0.01109	0.00011	0.01112	0.00011	0.01108
300	100	0.00003	0.0033	0.00003	0.00334	0.00003	0.00331	0.00003	0.00331
300	300	0.00001	0.0011	0.00001	0.00112	0.00001	0.00111	0.00001	0.00110

Results for sampling variances based on graphs

To illustrate more detailed patterns, U_M, B_M and V_M were plotted against sample sizes *K* and *I*, by the variance components σ_b^2, σ_w^2 and σ_e^2. Two graphs were created for U_M, where Figure 5.1 represents the cases with σ_b^2 = 1 and Figure 5.2 represents the cases with σ_b^2 = 100, using U_M as the vertical axis and combinations of *K* and *I* as the horizontal axis. Four lines were generated in each graph to represent U_M values by σ_w^2 and σ_e^2 . In the same format, Figures 5.3 and 5.4 were created for B_M and Figures 5.3 and 5.4 were created for V_M. Note that all the graphs in this study were created in this format, but the vertical scales may be different among graphs.

Besides the general patterns discussed based on the tabulation, these graphs show the level of impact from each simulation factor. For example, Figure 5.1 shows the U_M value at each level of sample sizes of the two sampling stages when $\sigma_b^2 = 1$

and the purple line with square markers in it represents the case with $\sigma_w^2 = 100$ and $\sigma_e^2 = 100$, etc. Figure 5.2 shows the parallel cases when $\sigma_b^2 = 100$. Every set of four points on the horizontal scale represent one level of the number of clusters (*K*) and, within the set, every point represent a level of cluster size (*I*). In both graphs, the overlap of the purple line with square markers with the green line with dot markers, and the red line with circle markers with the blue line with triangle markers, shows that σ_e^2 has no impact on U_M. By comparing the scale of the vertical axis of the two graphs, we can see that the positive impact from σ_b^2 to U_M is dominant, much larger than the positive impact from σ_w^2, which is shown by the differences between the green line with dot markers and red line with circle markers and between the purple line with square markers and blue line with triangle markers. According to the trend of each line, the sample sizes *K* and *I* have a negative impact on U_M. The difference of the shapes of the two graphs illustrates a large interaction effect between the number of clusters *K* and σ_b^2. That is, the negative effect of *K* on U_M is much larger when $\sigma_b^2 = 100$ than when $\sigma_b^2 = 1$.

Figures 5.3 and 5.4 for B_M tell a different story. The overlap of the red line with the green line, and of the purple line with the blue line, shows that σ_w^2 has no impact on B_M. The same shape and scale of the two graphs show that σ_b^2 has no impact on B_M. The differences between the green line and purple line and between the red line and blue line show the positive impact of σ_e^2 on B_M.

Figures 5.5 and 5.6 are for V_M, which is the summation of U_M and $1.1 \times B_M$. Thus, the impact from the simulation factors on V_M includes impacts from both sources. σ_b^2 still has the largest positive impact among the variance components, while the impact from σ_w^2 and σ_e^2 are at similar levels. The negative impact from the sample sizes on V_M is almost twice of the impact on U_M or B_M. Similar to the impact on U_M, the negative effect of K on V_M is much larger when $\sigma_b^2 = 100$ than when $\sigma_b^2 = 1$.

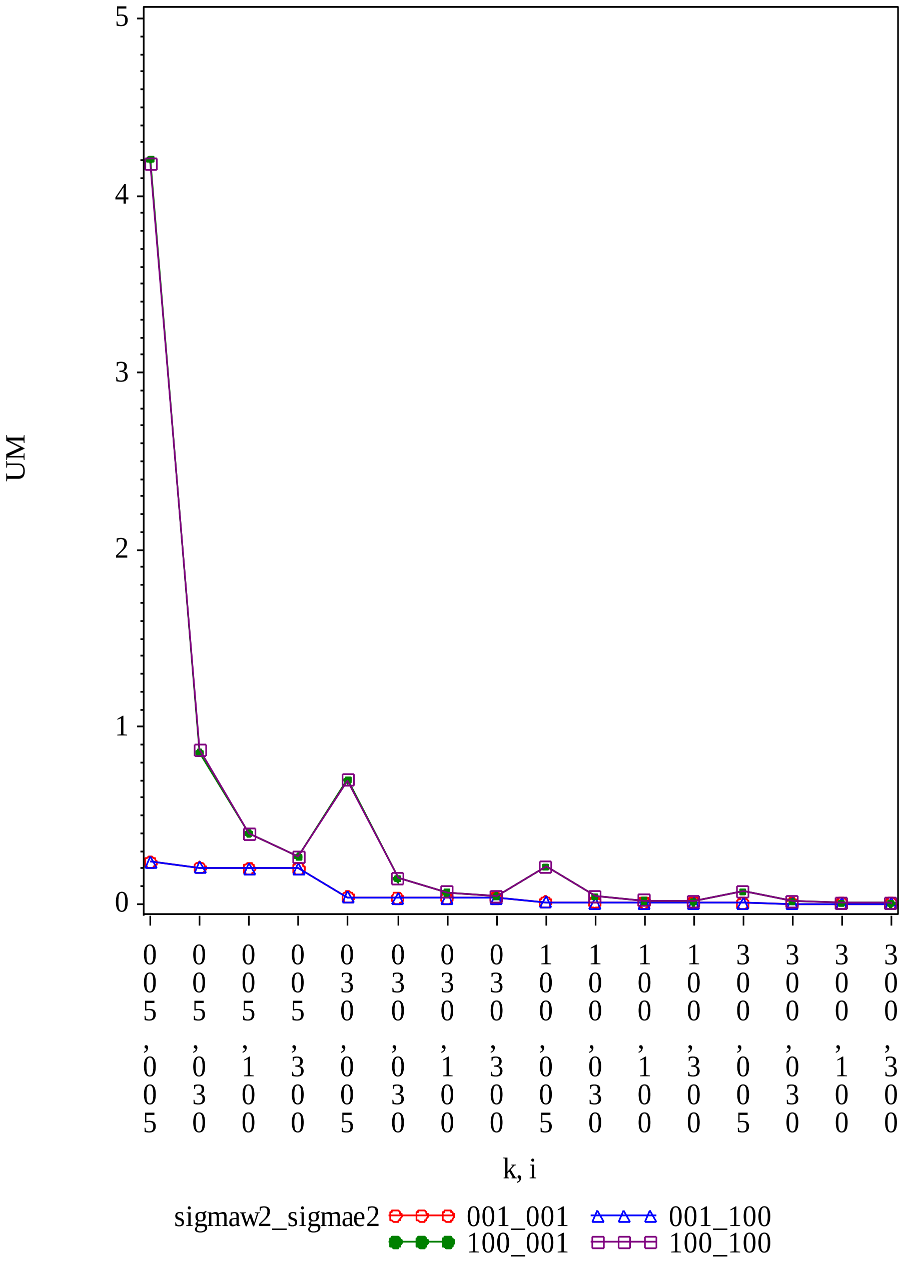

Figure 5.1 : U_M vs. sample sizes K and I by σ_w^2 and σ_e^2 (where $\sigma_b^2 = 1$, σ_w^2 = 1 or 100, and σ_e^2 = 1 or 100)

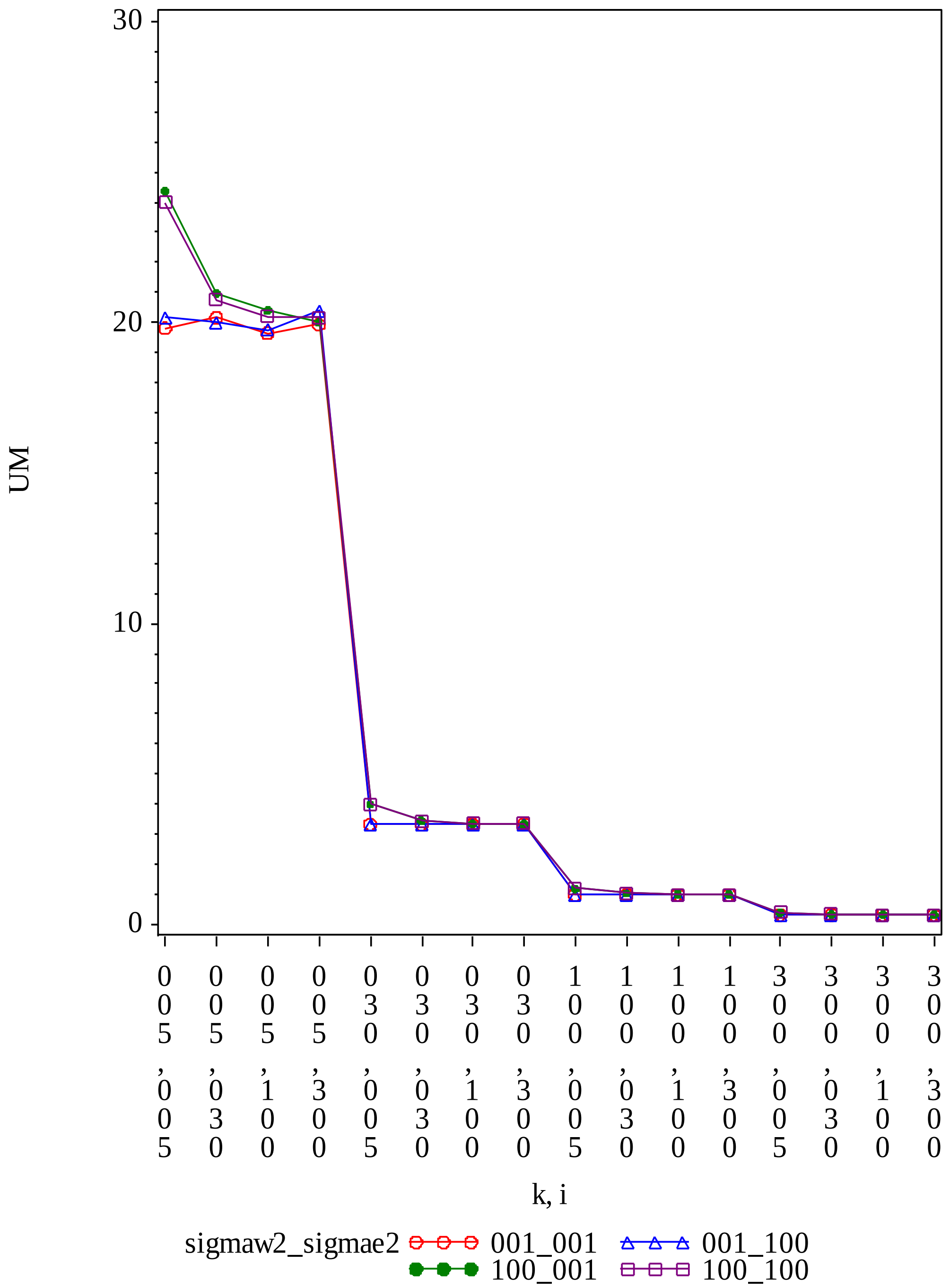

Figure 5.2 : U_M vs. sample sizes K and I by σ_w^2 and σ_e^2 (where $\sigma_b^2 = 100$, $\sigma_w^2 = 1$ or 100, and $\sigma_e^2 = 1$ or 100)

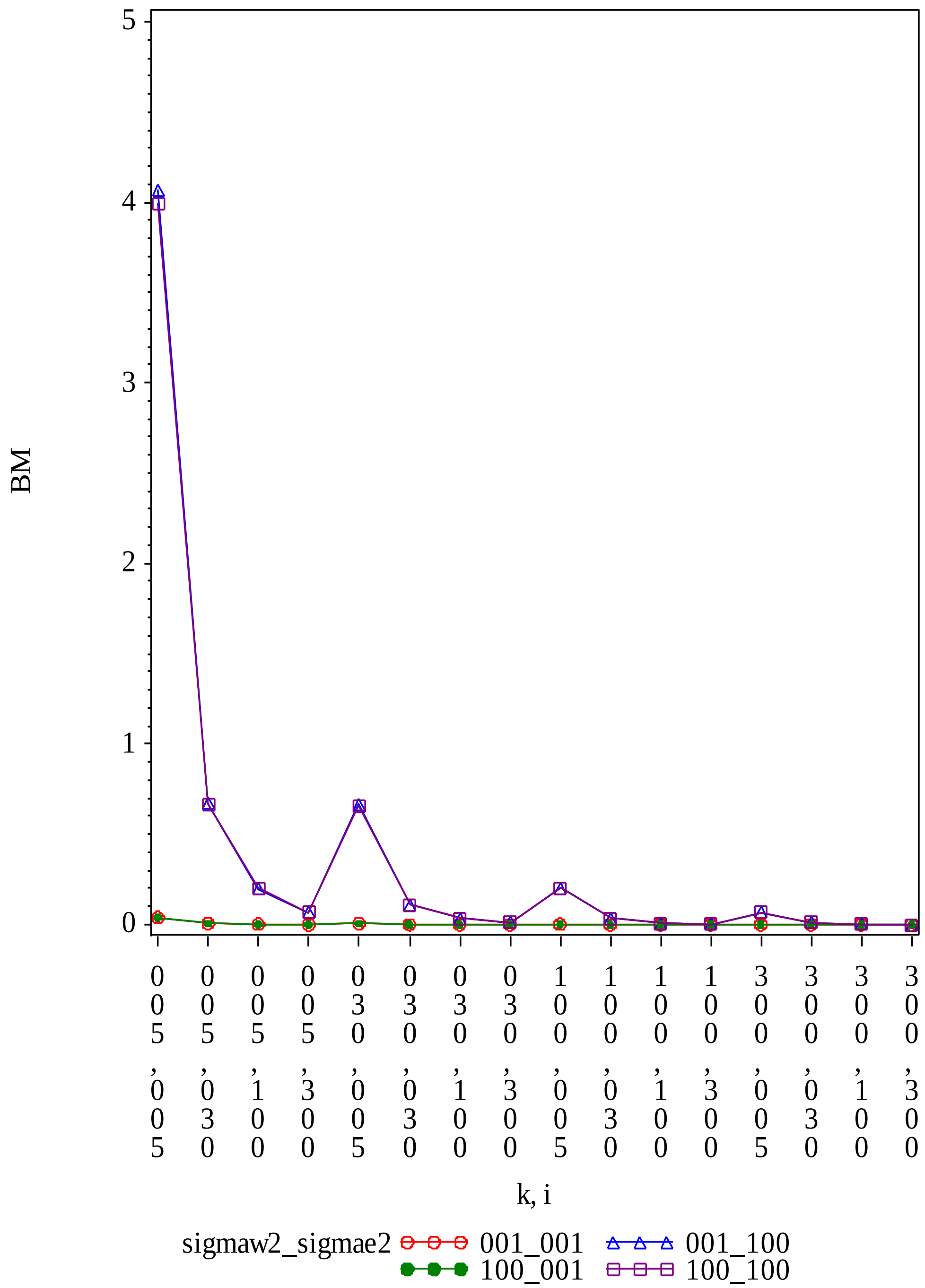

Figure 5.3 : B_M vs. sample sizes K and I by σ_w^2 and σ_e^2 (where $\sigma_b^2 = 1$, σ_w^2 = 1 or 100, and σ_e^2 = 1 or 100)

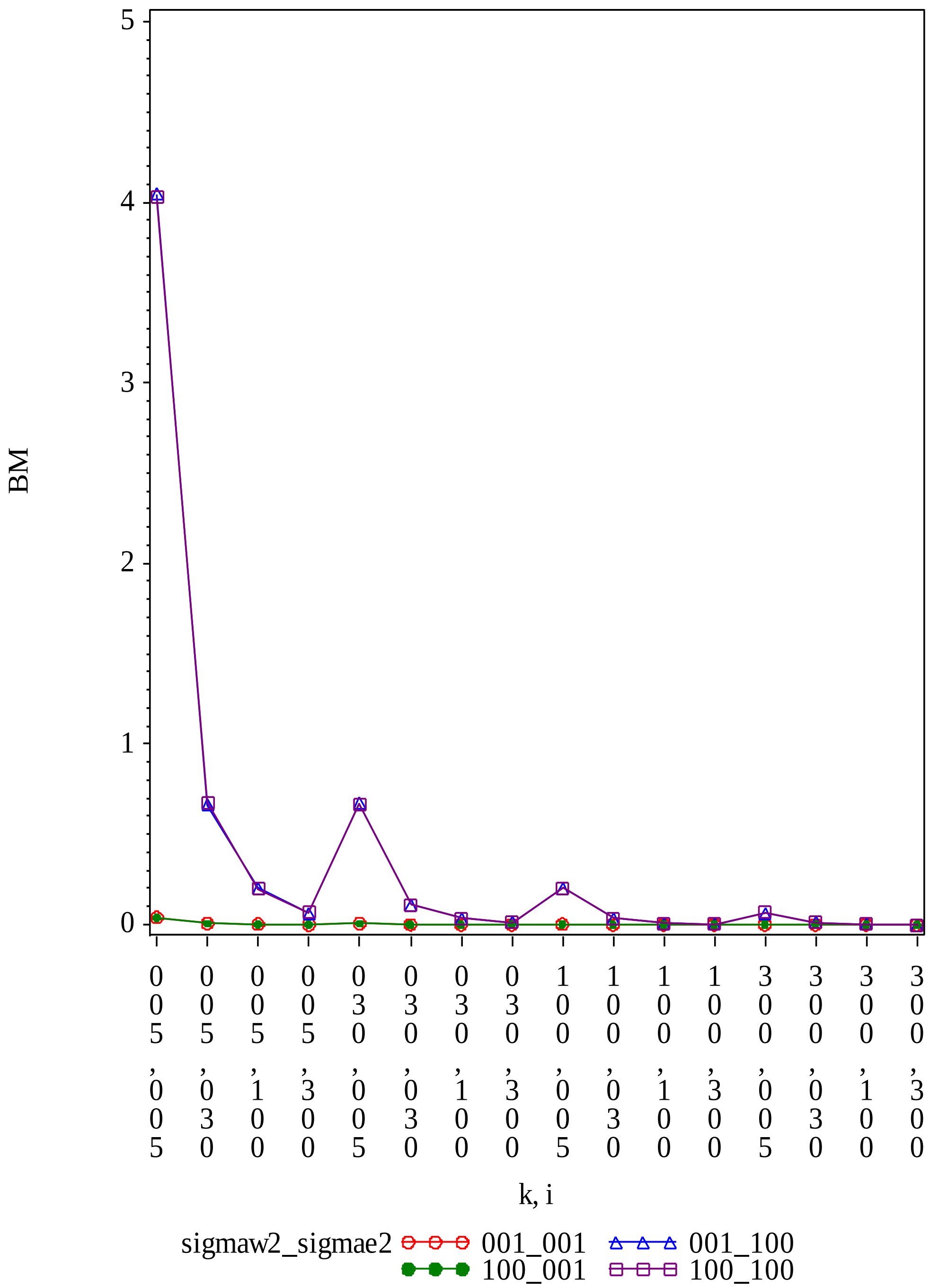

Figure 5.4: B_M vs. sample sizes K and I by σ_w^2 and σ_e^2 (where $\sigma_b^2 = 100$, σ_w^2 = 1 or 100, and σ_e^2 = 1 or 100)

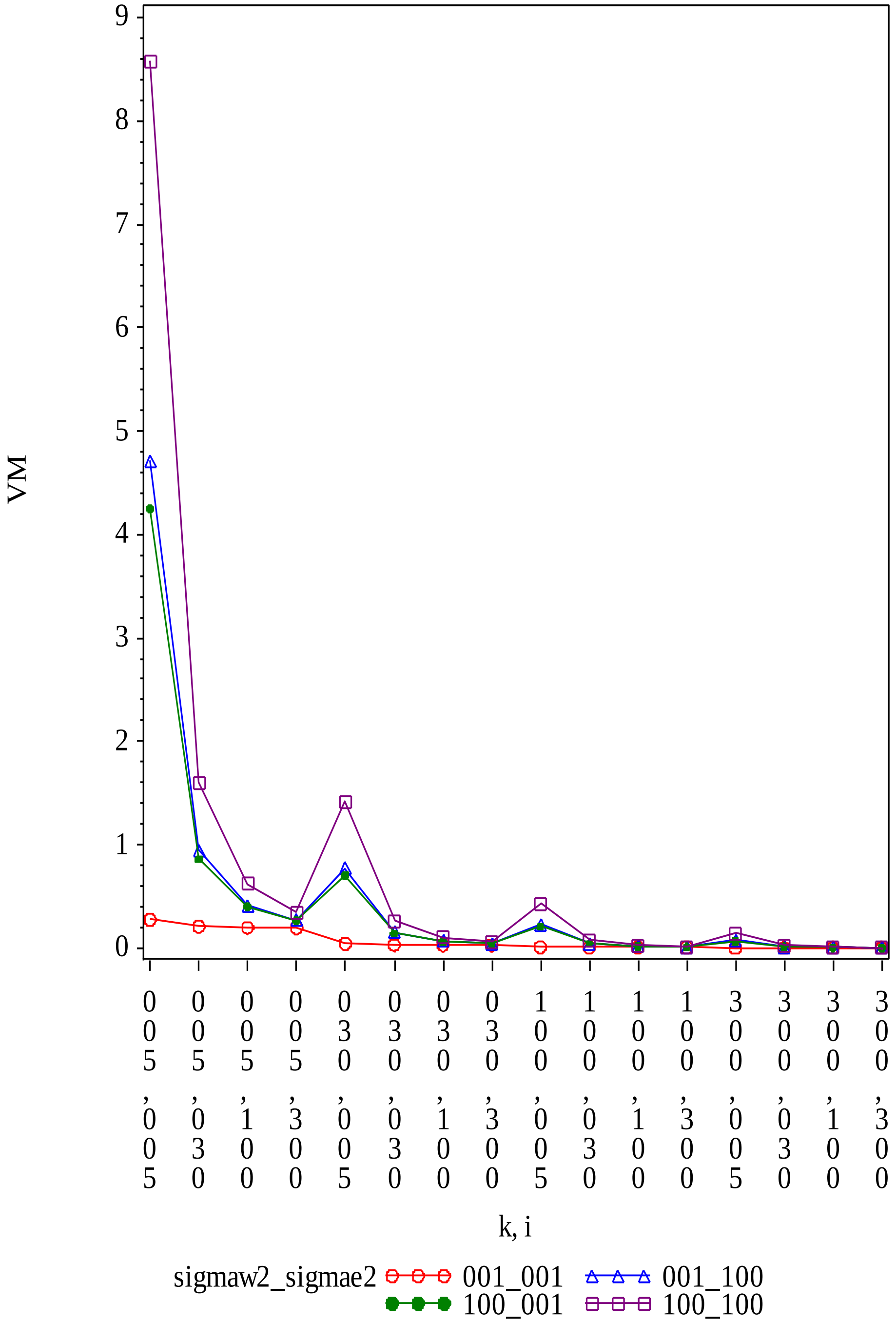

Figure 5.5: V_M vs. sample sizes *K* and *I* by σ_w^2 and σ_e^2 (where $\sigma_b^2 = 1$, σ_w^2 = 1 or 100, and σ_e^2 = 1 or 100)

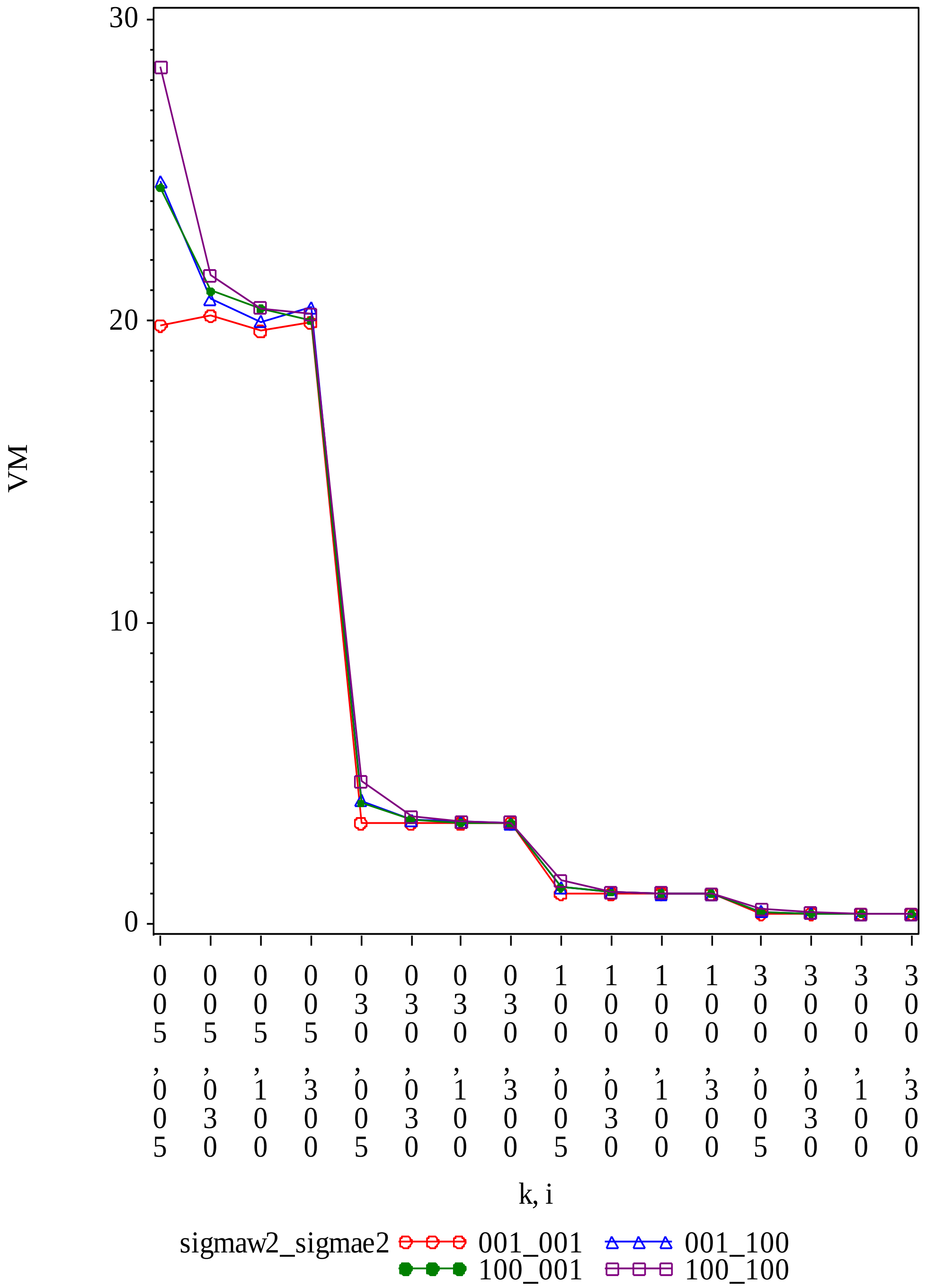

Figure 5.6: V_M vs. sample sizes *K* and *I* by σ_w^2 and σ_e^2 (where $\sigma_b^2 = 100$, σ_w^2 = 1 or 100, and σ_e^2 = 1 or 100)

Regression analyses on sampling variances

To quantify the effect of the simulation factors on the sampling variance of the plausible values, regression analyses were carried out for both simulation settings, the ratio of the variance components equal to 100 and 4. The model was fit for each of the outcome variables U_M, B_M and V_M. Moreover, the characteristic of V_M is illustrated by modeling the ratio of U_M over V_M.

Independent variables include σ_b^2, σ_w^2, σ_e^2, *K*, *I* and all the two-way interactions between these terms. As suggested by the shape of the graphs, the sample size variables in the quadratic form and their interactions with the variance components were added to the model. To improve the model fit, the cubic terms of the sample sizes and their interaction terms with the variance components were tested in the model and were kept when significant at the 0.05 level. The resulting model includes the following added terms: the quadratic terms of *K* and *I* and their interactions to the variance components, the cubic terms of *K* and *I* and the interaction between the terms for *K* and the between-cluster variance. Note that the baseline level of *K* and *I* were set to five and the baseline level for the variance components were zero. *K* and *I* were treated as continuous variables and the variance component variables were treated as binary variables.

Residual analysis for these models showed non-normal residuals with non-constant variance. Transformations in the forms of log, reciprocal, square root, had been considered and tested. However, transformations changed the relationship between the independent variables and the outcome variables dramatically, compared to the relationship shown in the graph. Although no transformation was implemented in these models, the models were still expected to make relatively sound inferences from the *F* test. As Lindman (1974) shows, the *F* statistic is quite robust against deviation from normal distribution and homogeneity of variances. Especially when the R^2 ‘s for the V_M and U_M are very close to 1, the impact of the non-normality in the distributions of residuals on the model statistics is limited, and is negligible on estimates of effects, which is the primary concern here. However, as a conservative approach, the *P*-values for the *F*-tests in the regression model should be considered as indicators of relative size of effects rather than taken at face.

The result of the residual analysis for cases with the ratio of the variance components equal to 100 is documented in Table 5.11, showing outlying residuals, skewness, and kurtosis for each model.

Table 5.11 Outlying residuals, skewness and kurtosis in the regression analysis, when variance components take values 1 and 100.

Outcome	Predictor						Distribution statistics	
	σ_b^2	σ_w^2	σ_e^2	K	I	Residual outliers	Skewness	Kurtosis
V_M	1	100	100	5	5	4.360	2.21	9.07
	100	100	100	5	5	4.314		
	100	1	100	5	5	2.077		
	1	1	100	5	5	2.021		
	100	100	1	5	5	1.959		
	1	100	1	5	5	1.684		
U_M	100	100	1	5	5	2.348	2.36	8.82
	1	100	100	5	5	2.076		
	1	100	1	5	5	2.071		
	100	100	100	5	5	1.999		
B_M	1	1	100	5	5	2.141	2.44	9.12
	100	1	100	5	5	2.116		
	100	100	100	5	5	2.105		
	1	100	100	5	5	2.077		
U_M/V_M	None						0.27	-0.39

Table 5.12 shows parameter estimates, P-values for the F-test and (semipartial) $\hat{\eta}^2$ values for each term, when the ratio of the variance components is 100. The (semipartial) $\hat{\eta}^2$statistic is defined as the proportion of total variation attributable to the predictor, partialling out other predictors from the total nonerror

variation for each predictor in the model, and provides a standard measure of the strength of the association between a predictor and the outcome variable. All variables including non-significant ones were kept in the model to show the importance of each predictor. Terms that are significant at the 0.1 level are in bold font in the table. Terms that explain larger variances of the outcome variables are highlighted.

The result of the regression analysis provides more detailed information about the relationship between the simulation factors and the sampling variances of the imputed scores. As discussed previously, the positive impact on V_M from σ_b^2 and σ_w^2 comes through U_M, while the positive impact from σ_e^2 comes through B_M. The parameter estimates show that the V_M estimate is larger by 19.85 (0.1985* 100), for the cases with $\sigma_b^2 = 100$ than for the cases with $\sigma_b^2 = 1$, at the baseline level of all other factors, that is, *K*=5, *I*=5, $\sigma_w^2 = 1$ and $\sigma_e^2 = 1$. The impacts from σ_w^2 are σ_e^2 are much smaller – the V_M estimate increases by 1.57 when σ_w^2 changes from 1 to 100 and increases by 1.70 when σ_e^2 changes from 1 to 100. The negative significant interaction terms $\sigma_b^2 * \sigma_w^2$ and $\sigma_b^2 * \sigma_e^2$ show that the impact from σ_b^2 is less when σ_w^2 or σ_e^2 is 100.

According to the graph, the impacts from the sample sizes are in a curved shape, which is confirmed by the parameter estimates in the model and characterized by the main effect, the quadratic term and the cubic term of *K* and *I*. In specific, the parameter estimates of the main effects shows, in average, both *K* and *I* have negative impacts on V_M, as well as on U_M and B_M. Significant positive quadratic terms and smaller negative coefficients for the cubic terms for *K* and *I* illustrate that the slope of the curve gets flatter for larger sample sizes. In addition, the positive significant interaction terms between *K* and *I* show that the (negative) impact from *K* (or *I*) on V_M, U_M and B_M is smaller when *I* (or *K*) is higher.

In terms of the significance of the interaction terms between the variance components and the sample sizes, similar patterns are shown as those of the corresponding main effect terms for the variance components and the sample sizes. For σ_b^2, the interaction terms with K, K^2,K^3 are significant in making inferences on both V_M and U_M; for σ_w^2, the interaction terms with K, K^2, I and I^2 are significant for both V_M and U_M; and for σ_e^2, the interaction terms with K, K^2, I and I^2 are significant in making inferences on both V_M and B_M. The signs of these interaction terms are the same as the corresponding main effect terms for the sample sizes, K, K^2,K^3, I and I^2, showing that the impact from the sample sizes on the sampling variance is larger when the variance components are larger (100). Note that σ_b^2 does not have a significant interaction with *I*.

No interactions between the variance components are significant.

Table 5.12 reports $\hat{\eta}^2$ based on the Type III sum of square (SS). By examining this statistic, we can identify predictors with major impact on the outcome variables, after partialling out the effect from other predictors. For V_M and U_M, σ_b^2 has the largest impact (over 26% of the total variance), and along with its interactions with *K*, over 60% of the variance explained by the model (R^2=0.9848 or 0.9946) or the total variance is accounted for. For B_M, σ_e^2 has the largest impact (over 17% of the total variance), and along with its interactions with *K* and *I*, over 70% of the variance explained by the model (R^2=0.5356) or over 38% of the total variance is accounted for.

Table 5.12 Parameter estimates, *P*-values of *F*-tests and semipartial $\hat{\eta}^2$'s for the regression models on U_M , B_M and V_M , when the ratio of the variance components is 100

Outcome Variables	V_M (R^2=0.9848)			U_M (R^2=0.9946)			B_M (R^2=0.5356)		
Parameter	Estimate	*P*-value	$\hat{\eta}^2$	Estimate	*P*-value	$\hat{\eta}^2$	Estimate	*P*-value	$\hat{\eta}^2$
σ_b^2	**0.19851**	**<0.001**	**0.2613**	**0.19852**	**<0.001**	**0.2812**	-6.45E-06	0.998	<0.0001
σ_w^2	**0.01569**	**0.001**	**0.0019**	**0.01576**	**<0.001**	**0.0021**	-6.59E-05	0.979	<0.0001
σ_e^2	**0.01698**	**<0.001**	**0.0022**	0.00014	0.956	<0.0001	**0.01531**	**<0.001**	**0.1740**
K	**-0.05323**	**0.009**	**0.0011**	**-0.02953**	**0.012**	**0.0004**	**0.059**	**0.0167**	**0.059**
K^2	**0.00066**	**0.010**	**0.0010**	**0.00036**	**0.014**	**0.0003**	**0.057**	**0.0171**	**0.057**
K^3	**-1.68E-06**	**0.010**	**0.0010**	**-9.12E-07**	**0.015**	**0.0003**	**0.056**	**0.0172**	**0.056**
I	**-0.04407**	**0.004**	**0.0013**	**-0.02044**	**0.020**	**0.0003**	**0.013**	**0.0296**	**0.013**
I^2	**0.00056**	**0.002**	**0.0015**	**0.00026**	**0.014**	**0.0003**	**0.008**	**0.0338**	**0.008**
I^3	**-1.42E-06**	**0.002**	**0.0015**	**-6.57E-07**	**0.013**	**0.0003**	**0.008**	**0.0342**	**0.008**
$K*I$	**0.00002**	**0.003**	**0.0014**	**9.12E-06**	**0.014**	**0.0003**	**9.22E-06**	**0.012**	**0.0304**
$\sigma_b^2 * K$	**-0.00880**	**<0.001**	**0.1562**	**-0.00880**	**<0.001**	**0.1681**	4.51E-07	0.998	<0.0001
$\sigma_b^2 * K^2$	**0.00009**	**<0.001**	**0.1026**	**0.00009**	**<0.001**	**0.1104**	-6.32E-09	0.997	<0.0001
$\sigma_b^2 * K^3$	**-2.21E-07**	**<0.001**	**0.0895**	**-2.21E-07**	**<0.001**	**0.0963**	1.62E-11	0.997	<0.0001
$\sigma_b^2 * I$	-7.65E-06	0.914	<0.0001	-7.51E-06	0.854	<0.0001	-1.32E-07	0.997	<0.0001
$\sigma_b^2 * I^2$	2.49E-08	0.912	<0.0001	2.45E-08	0.850	<0.0001	3.54E-10	0.998	<0.0001
$\sigma_w^2 * K$	**-0.00015**	**0.031**	**0.0007**	**-0.00015**	**<0.001**	**0.0008**	6.54E-07	0.987	<0.0001
$\sigma_w^2 * K^2$	**4.11E-07**	**0.069**	**0.0005**	**4.13E-07**	**0.002**	**0.0006**	-1.74E-09	0.989	<0.0001
$\sigma_w^2 * I$	**-0.00014**	**0.050**	**0.0006**	**-0.00014**	**0.001**	**0.0006**	8.13E-07	0.984	<0.0001
$\sigma_w^2 * I^2$	**3.65E-07**	**0.105**	**0.0004**	**3.67E-07**	**0.005**	**0.0004**	-2.23E-09	0.986	<0.0001
$\sigma_e^2 * K$	**-0.00016**	**0.023**	**0.0008**	-9.03E-07	0.982	<0.0001	**-0.00015**	**<0.001**	**0.0623**
$\sigma_e^2 * K^2$	**4.32E-07**	**0.056**	**0.0006**	2.39E-09	0.985	<0.0001	**3.90E-07**	**0.003**	**0.0439**
$\sigma_e^2 * I$	**-0.00016**	**0.024**	**0.0008**	-7.31E-07	0.986	<0.0001	**-0.00015**	**<0.001**	**0.0622**
$\sigma_e^2 * I^2$	**4.37E-07**	**0.053**	**0.0006**	8.06E-09	0.950	<0.0001	**3.90E-07**	**0.003**	**0.0439**
$\sigma_b^2 * \sigma_w^2$	4.04E-06	0.906	<0.0001	3.82E-06	0.846	<0.0001	2.02E-07	0.992	<0.0001
$\sigma_b^2 * \sigma_e^2$	5.51E-07	0.987	<0.0001	4.93E-07	0.980	<0.0001	5.23E-08	0.998	<0.0001
$\sigma_w^2 * \sigma_e^2$	-4.86E-06	0.887	<0.0001	-4.51E-06	0.818	<0.0001	-3.16E-07	0.987	<0.0001

Analysis of ratio variables – graphs and regression

To fully examine the characteristics of sampling variance of the imputed score based on MI, the ratios of the within imputation variance U_M to the total sampling variance V_M were derived and their relationship to the simulation variables were analyzed. This ratio shows the proportion of the sampling variance accounted for by single imputation rather than multiple imputation and the supplement of this ratio reflects what proportion of the missing information is due to not observing the true score directly.

Graphs were generated by plotting the ratio of U_M to V_M against sample sizes *K* and *I*, by the variance components σ_b^2 , σ_w^2 and σ_e^2 . Two graphs were created, where σ_b^2 =1 in the first graph and σ_b^2 =100 in the second, using U_M/V_M as the vertical axis and combinations of *K* and *I* as the horizontal axis. Four lines were generated to represent combinations of σ_w^2 and σ_e^2 According to the first graph, the ratio is positively related to cluster size *I* and σ_w^2 and negatively related to σ_e^2 . The number of clusters *K* has no impact on the ratio. Compared to the second graph, the plots in the first graph change within a larger range as seen in the scale of the vertical axis. Thus, σ_b^2 has a negative impact to the ratio. Moreover, the slope of the lines for each set of *K* values at the same level of *I* is different among all combinations of variance components. This illustrates the interaction between *K* and the variance components. The discussion of the regression analysis in the next paragraph includes more details.

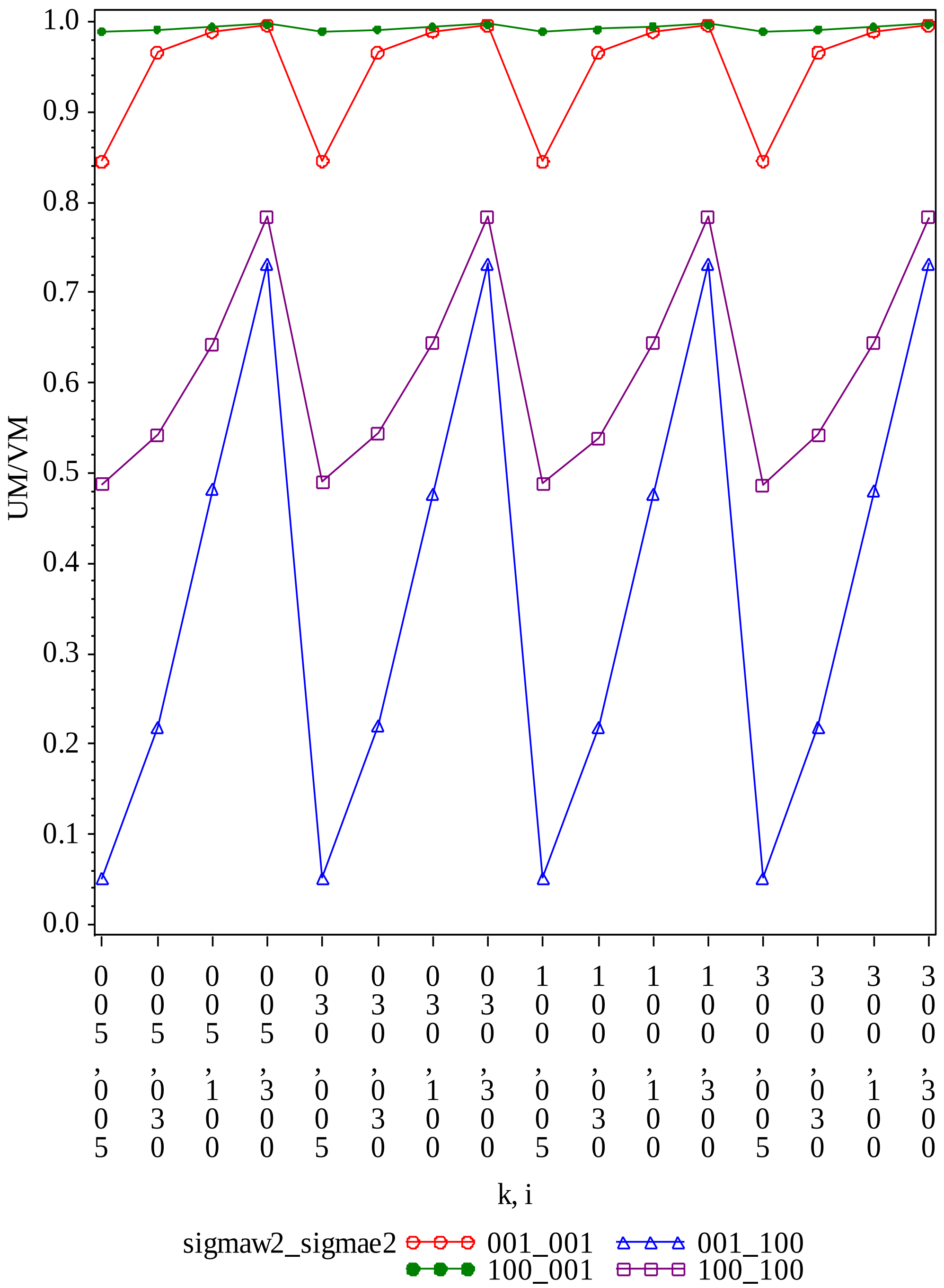

Figure 5.7: U_M / V_M vs. sample size K and I by σ_w^2 and σ_e^2 (where σ_b^2 =1, σ_w^2 = 1 or 100, and σ_e^2 = 1 or 100)

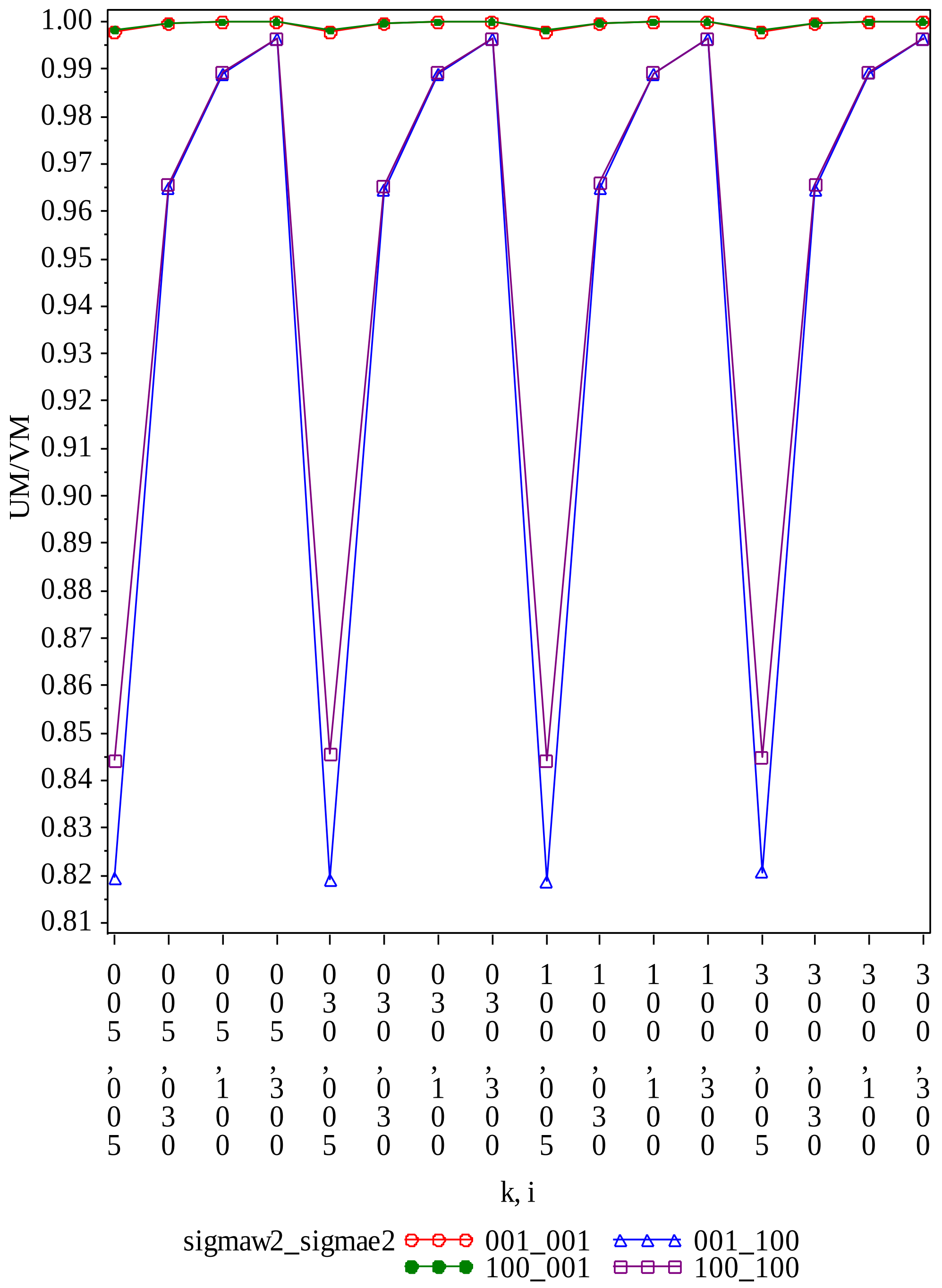

Figure 5.8: U_M / V_M vs. sample size K and I by σ_w^2 and σ_e^2 (where σ_b^2 =100, σ_w^2 = 1 or 100, and σ_e^2 = 1 or 100)

Regression analyses were carried out for the ratio variable and the results are shown in table 5.13. Terms that are significant at the 0.1 level are in bold font in the table. Terms that explain larger variances of the outcome variables are highlighted.

All three main effects of the variance component variables and their interactions with each other are significant for the ratio variable. For example, when σ_b^2 or σ_w^2 are larger or when σ_e^2 is smaller, holding other variables at the baseline level, the ratio variable has higher values. The significant interaction terms between variance components show that the effect of one variance component is different when the other components are not at the baseline level. In other words, when both factors change from 1 to 100, besides the main effects, there is a further change on the outcome variable. For example, when both σ_b^2 and σ_e^2 change from 1 to 100, the estimated ratio of U_M to V_M changes by 0.0000425*100, besides the positive main effect (0.0018440*100) from σ_b^2 and the negative main effect (-0.0065681 *100) from σ_e^2.

In terms of the effect of the sample size, the main effect, the quadratic term, and the cubic term of cluster size *I* are significant, as are the interaction terms with variance components $\sigma_b^2 * I$, $\sigma_w^2 * I$, $\sigma_w^2 * I^2$, $\sigma_e^2 * I$, and $\sigma_e^2 * I^2$. Note that no term involving the number of clusters *K* is significant.

Semipartial $\hat{\eta}^2$ in the table shows that the major impact to the ratio is from σ_e^2 (27.36%) and $\sigma_b^2 * \sigma_e^2$ (19.19%).

Table 5.13 Parameter estimates, P-values of F-tests and semipartial $\hat{\eta}^2$'s for the regression model on the ratio variable U_M/V_M when the ratio of the variance components is 100.

Outcome Variables	U_M/V_M (R^2=0.9579)		
Parameter	Estimate	P-value	$\hat{\eta}^2$
σ_b^2	**1.84E-03**	**<0.001**	**0.0184**
σ_w^2	**1.58E-03**	**<0.001**	**0.0158**
σ_e^2	**-6.57E-03**	**<0.001**	**0.2736**
K	2.81E-05	0.981	<0.0001
K^2	-3.95E-07	0.979	<0.0001
K^3	1.03E-09	0.978	<0.0001
I	**0.00403**	**<0.001**	**0.0089**
I^2	**-3.33E-05**	**0.002**	**0.0043**
I^3	**7.51E-08**	**0.005**	**0.0034**
$K*I$	-1.69E-10	1.000	<0.0001
$\sigma_b^2 * K$	-2.41E-07	0.988	<0.0001
$\sigma_b^2 * K^2$	3.64E-09	0.986	<0.0001
$\sigma_b^2 * K^3$	-9.56E-12	0.986	<0.0001
$\sigma_b^2 * I$	**-1.32E-05**	**0.002**	**0.0043**
$\sigma_b^2 * I^2$	2.09E-08	0.111	0.0011
$\sigma_w^2 * K$	-1.80E-10	1.000	<0.0001
$\sigma_w^2 * K^2$	-4.62E-11	0.997	<0.0001
$\sigma_w^2 * I$	**-1.40E-05**	**0.001**	**0.0048**
$\sigma_w^2 * I^2$	**3.30E-08**	**0.013**	**0.0027**
$\sigma_e^2 * K$	-4.32E-08	0.992	<0.0001
$\sigma_e^2 * K^2$	1.35E-10	0.992	<0.0001
$\sigma_e^2 * I$	**2.35E-05**	**<0.001**	**0.0138**
$\sigma_e^2 * I^2$	**-4.87E-08**	**<0.001**	**0.0058**
$\sigma_b^2 * \sigma_w^2$	**-1.41E-05**	**<0.001**	**0.0210**
$\sigma_b^2 * \sigma_e^2$	**4.25E-05**	**<0.001**	**0.1919**
$\sigma_w^2 * \sigma_e^2$	**1.03E-05**	**<0.001**	**0.0113**

5.5.2.2 The ratio of the variance components is 4

In the analysis in this section, the variance components (σ_b^2, σ_w^2 and σ_e^2) take values 1 and 4. The structure of this section is similar to the previous section 5.5.2.1.

Tabulation of sampling variances by simulation factors

Tables 5.14 – 5.16 are parallel to tables 5.8 – 5.10, showing statistics V_M, U_M and B_M, respectively, by the simulation factors. Except that the sampling variances are in a smaller scale, similar patterns are found in these tables.

Table 5.14 V_M for the plausible values by the simulation factors.

Sample size		Variance Components $\sigma_b^2/\sigma_w^2/\sigma_e^2$							
K	I	1/1/1	1/1/4	1/4/1	1/4/4	4/1/1	4/1/4	4/4/1	4/4/4
5	5	0.284	0.419	0.407	0.531	0.881	1.008	1.006	1.135
5	30	0.212	0.236	0.233	0.256	0.817	0.831	0.830	0.859
5	100	0.206	0.212	0.208	0.218	0.805	0.805	0.807	0.811
5	300	0.204	0.204	0.204	0.207	0.792	0.799	0.806	0.808
30	5	0.048	0.069	0.067	0.089	0.147	0.169	0.167	0.189
30	30	0.036	0.039	0.039	0.043	0.136	0.140	0.140	0.143
30	100	0.034	0.035	0.035	0.036	0.134	0.135	0.136	0.136
30	300	0.033	0.034	0.034	0.034	0.134	0.134	0.135	0.134
100	5	0.014	0.021	0.020	0.027	0.044	0.051	0.050	0.057
100	30	0.011	0.012	0.012	0.013	0.041	0.042	0.042	0.043
100	100	0.010	0.011	0.010	0.011	0.040	0.041	0.041	0.041
100	300	0.010	0.010	0.010	0.010	0.040	0.040	0.040	0.040
300	5	0.005	0.007	0.007	0.009	0.015	0.017	0.017	0.019
300	30	0.004	0.004	0.004	0.004	0.014	0.014	0.014	0.014
300	100	0.003	0.004	0.004	0.004	0.013	0.014	0.013	0.014
300	300	0.003	0.003	0.003	0.003	0.013	0.013	0.013	0.013

Table 5.15 U_M for the plausible values by the simulation factors.

Sample size		Variance components $\sigma_b^2/\sigma_w^2/\sigma_e^2$							
K	*I*	1/1/1	1/1/4	1/4/1	1/4/4	4/1/1	4/1/4	4/4/1	4/4/4
5	5	0.241	0.243	0.363	0.357	0.838	0.832	0.962	0.959
5	30	0.205	0.207	0.226	0.227	0.810	0.802	0.822	0.830
5	100	0.204	0.203	0.206	0.209	0.803	0.797	0.805	0.802
5	300	0.203	0.201	0.203	0.204	0.791	0.796	0.806	0.805
30	5	0.040	0.040	0.060	0.060	0.140	0.140	0.159	0.159
30	30	0.035	0.034	0.037	0.038	0.135	0.135	0.138	0.138
30	100	0.034	0.034	0.035	0.035	0.134	0.133	0.135	0.135
30	300	0.033	0.034	0.034	0.034	0.134	0.134	0.135	0.133
100	5	0.012	0.012	0.018	0.018	0.042	0.042	0.048	0.048
100	30	0.010	0.010	0.011	0.011	0.040	0.040	0.041	0.041
100	100	0.010	0.010	0.010	0.010	0.040	0.040	0.040	0.040
100	300	0.010	0.010	0.010	0.010	0.040	0.040	0.040	0.040
300	5	0.004	0.004	0.006	0.006	0.014	0.014	0.016	0.016
300	30	0.003	0.003	0.004	0.004	0.013	0.013	0.014	0.014
300	100	0.003	0.003	0.003	0.003	0.013	0.013	0.013	0.013
300	300	0.003	0.003	0.003	0.003	0.013	0.013	0.013	0.013

Table 5.16 B_M for the plausible values by the simulation factors.

Sample size		Variance components $\sigma_b^2/\sigma_w^2/\sigma_e^2$							
K	*I*	1/1/1	1/1/4	1/4/1	1/4/4	4/1/1	4/1/4	4/4/1	4/4/4
5	5	0.03953	0.16049	0.04000	0.15853	0.03967	0.16043	0.03976	0.16006
5	30	0.00668	0.02674	0.00670	0.02689	0.00665	0.02676	0.00673	0.02654
5	100	0.00200	0.00816	0.00198	0.00804	0.00200	0.00805	0.00200	0.00798
5	300	0.00067	0.00267	0.00067	0.00269	0.00067	0.00267	0.00068	0.00269
30	5	0.00671	0.02661	0.00664	0.02651	0.00669	0.02654	0.00665	0.02678
30	30	0.00112	0.00447	0.00111	0.00445	0.00110	0.00442	0.00111	0.00445
30	100	0.00033	0.00132	0.00033	0.00133	0.00034	0.00134	0.00033	0.00134
30	300	0.00011	0.00045	0.00011	0.00045	0.00011	0.00044	0.00011	0.00045
100	5	0.00198	0.00796	0.00201	0.00806	0.00202	0.00809	0.00197	0.00808
100	30	0.00033	0.00134	0.00033	0.00133	0.00033	0.00133	0.00034	0.00134
100	100	0.00010	0.00040	0.00010	0.00040	0.00010	0.00040	0.00010	0.00040
100	300	0.00003	0.00013	0.00003	0.00013	0.00003	0.00013	0.00003	0.00013
300	5	0.00067	0.00268	0.00067	0.00266	0.00067	0.00268	0.00067	0.00267
300	30	0.00011	0.00044	0.00011	0.00045	0.00011	0.00044	0.00011	0.00045
300	100	0.00003	0.00013	0.00003	0.00013	0.00003	0.00013	0.00003	0.00013
300	300	0.00001	0.00005	0.00001	0.00004	0.00001	0.00004	0.00001	0.00004

Results for sampling variances based on graphs

The shapes of the figures are similar to the figures where the ratio of the variance components is 100. The patterns discovered and discussed previously are also observed in this set of figures. Detailed examination is included in the regression analysis.

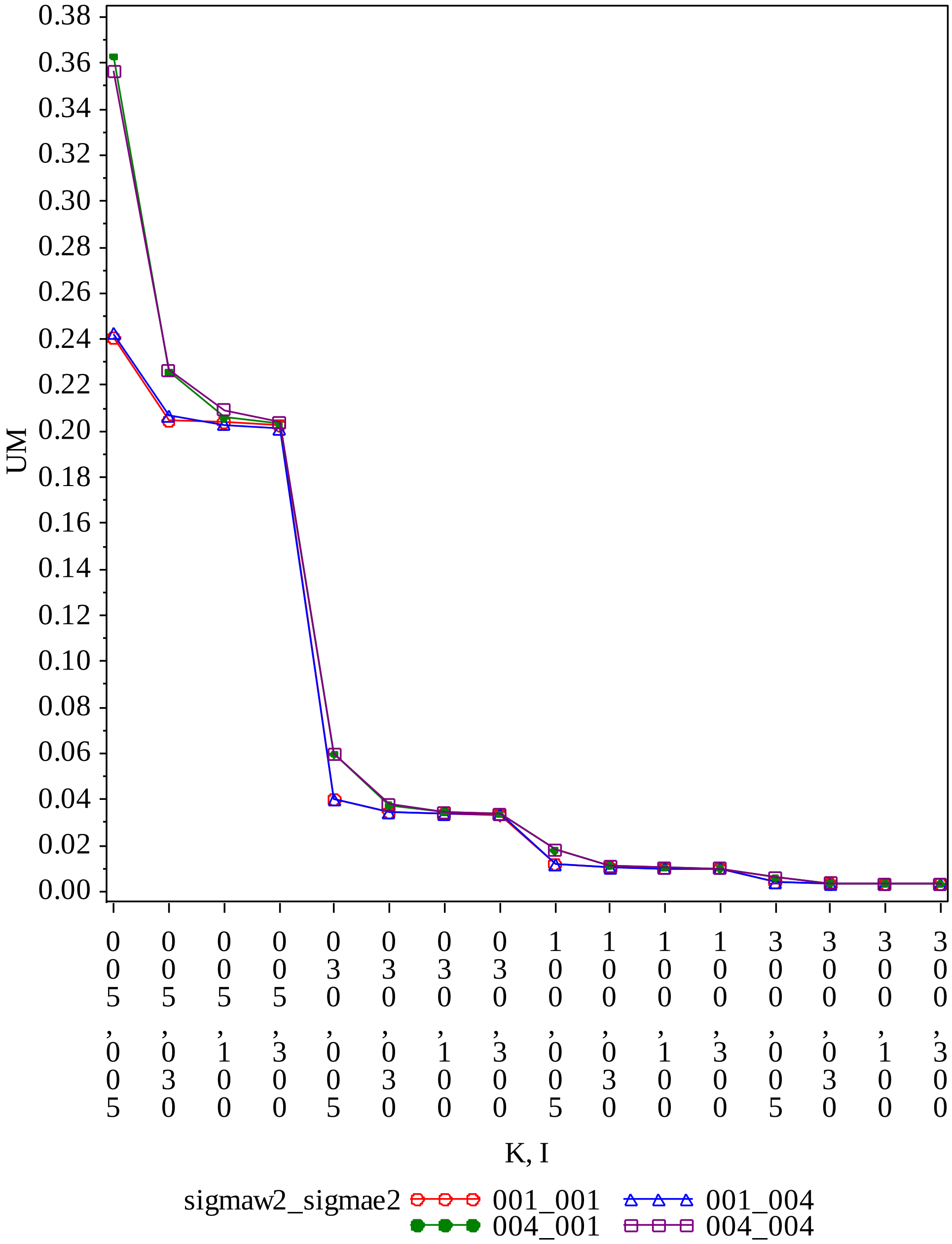

Figure 5.9: U_M vs. sample size K and I by σ_w^2 and σ_e^2 (where σ_b^2 =1, σ_w^2 = 1 or 4, and σ_e^2 = 1 or 4)

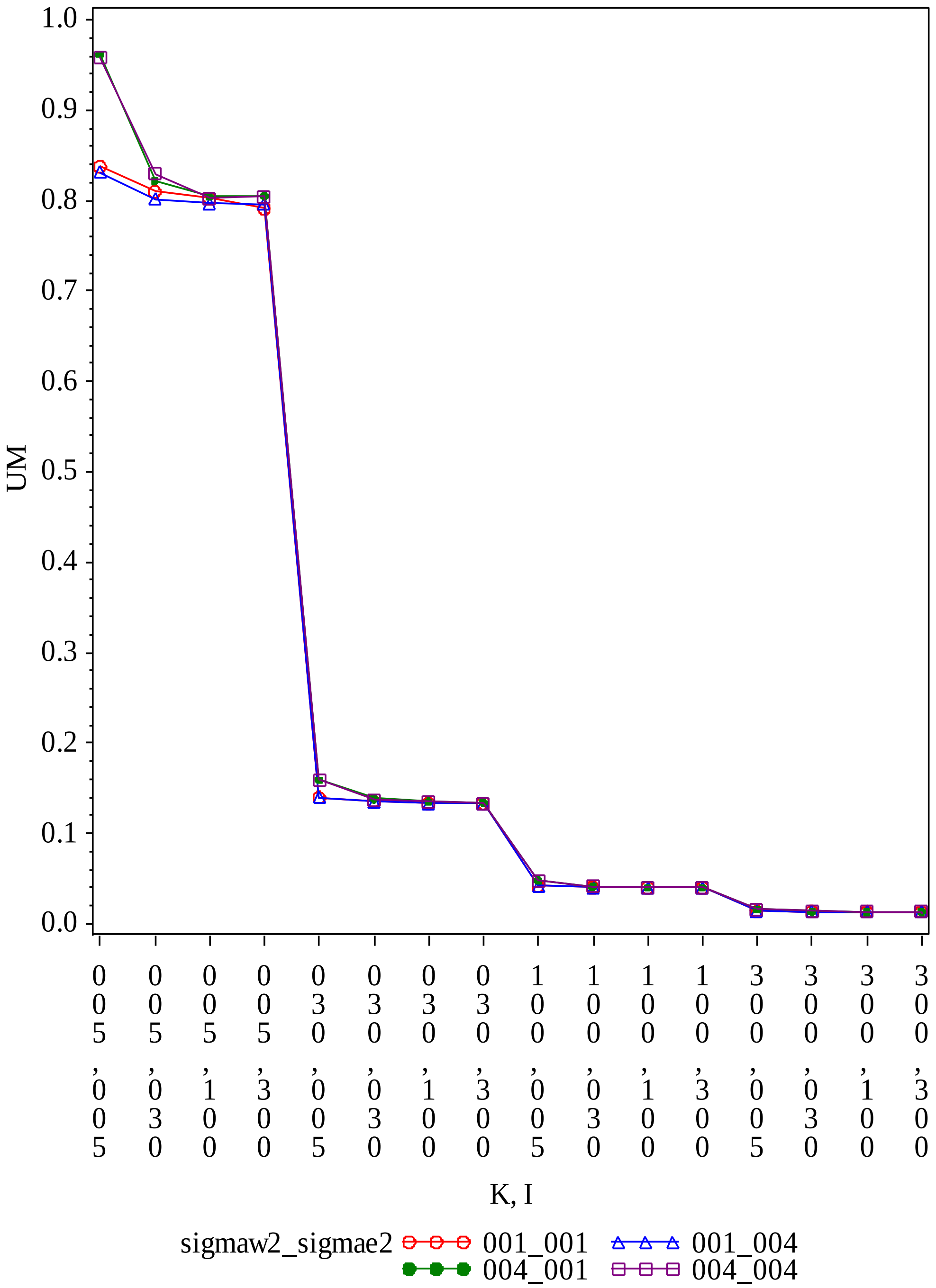

Figure 5.10: U_M vs. sample size *K* and *I* by σ_w^2 and σ_e^2 (where σ_b^2 =4, σ_w^2 = 1 or 4, and σ_e^2 = 1 or 4)

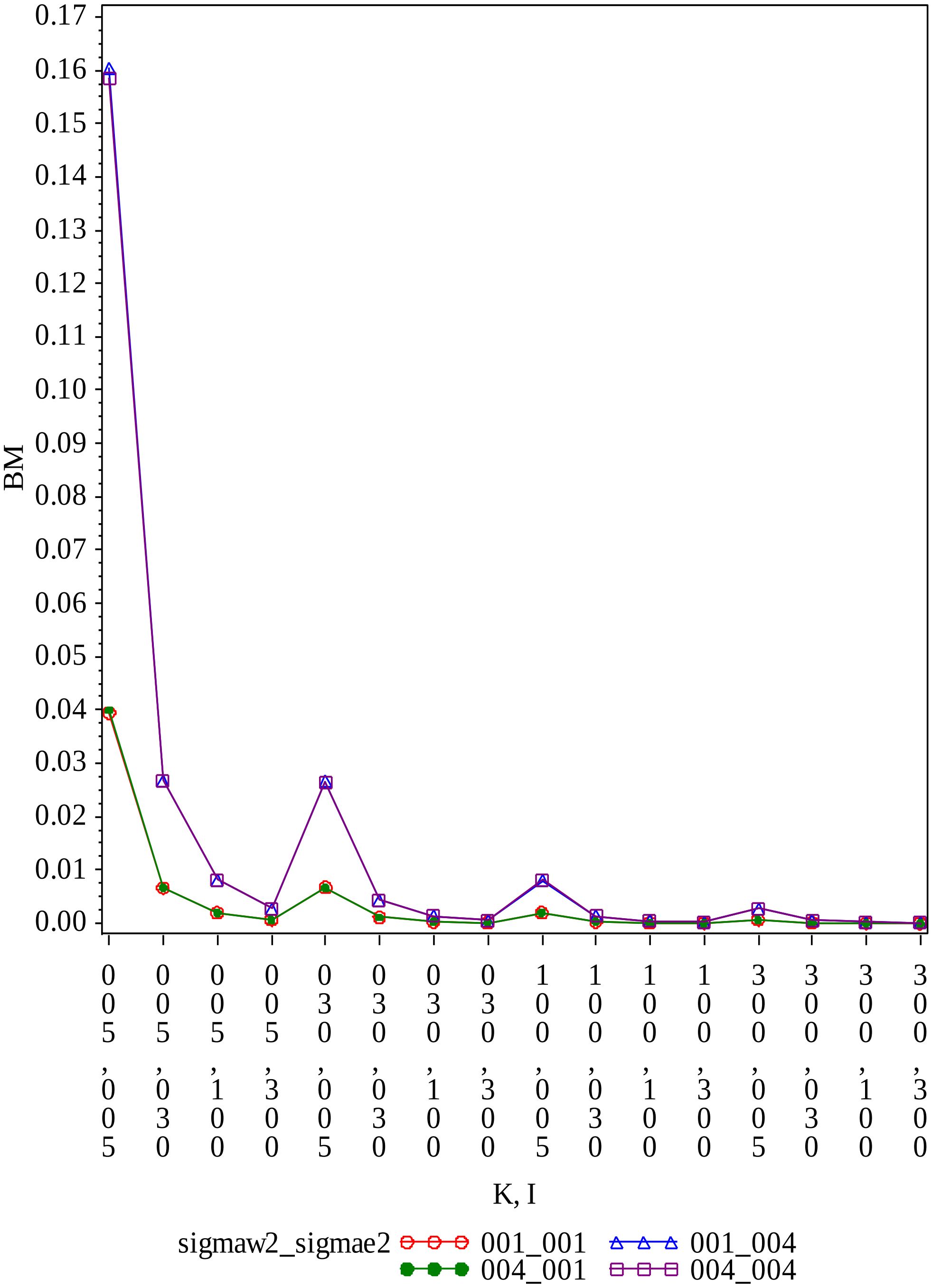

Figure 5.11: B_M vs. sample size K and I by σ_w^2 and σ_e^2 (where σ_b^2 =1, σ_w^2 = 1 or 4, and σ_e^2 = 1 or 4)

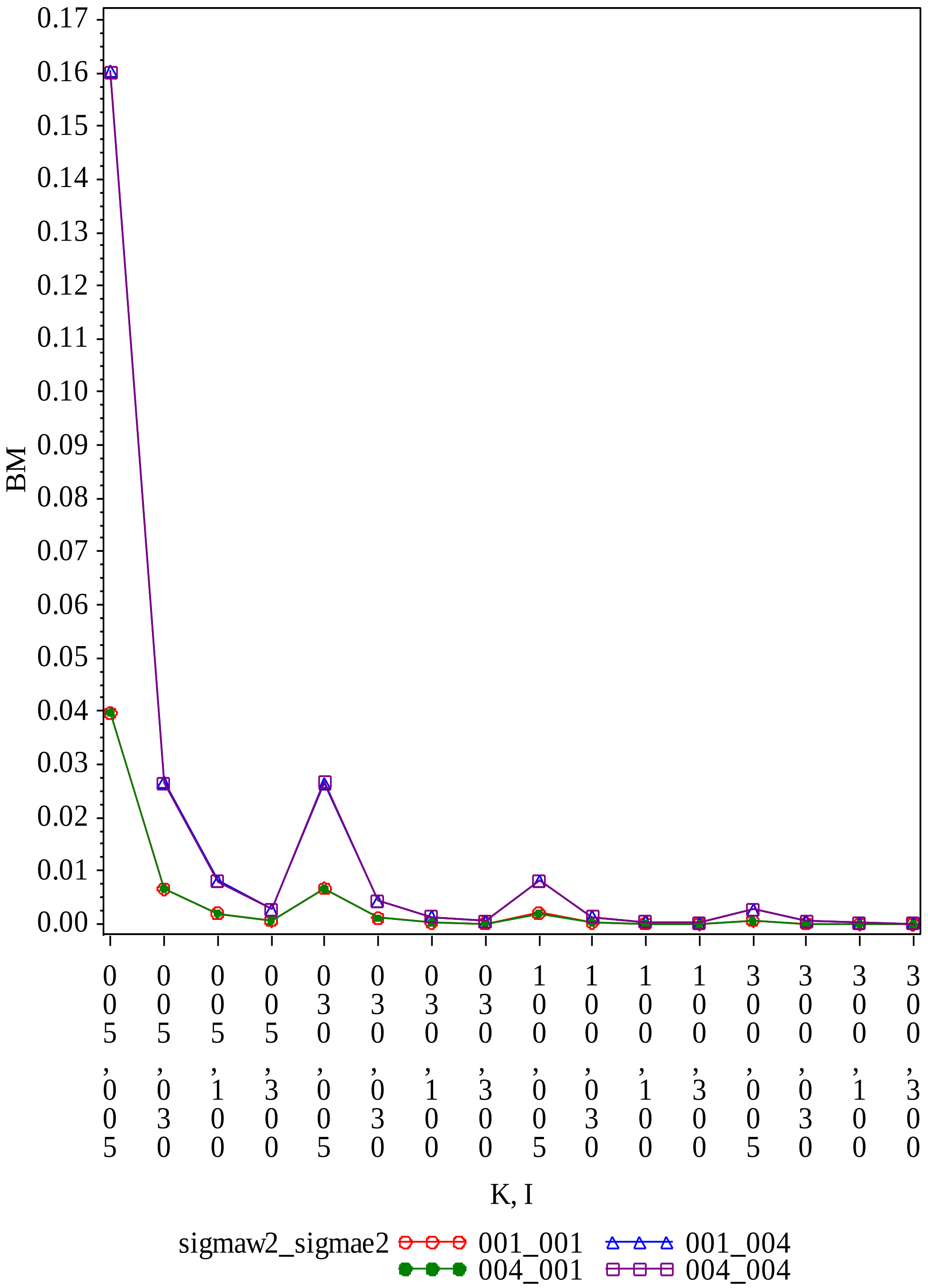

Figure 5.12: B_M vs. sample size K and I by σ_w^2 and σ_e^2 (where σ_b^2 =4, σ_w^2 = 1 or 4, and σ_e^2 = 1 or 4)

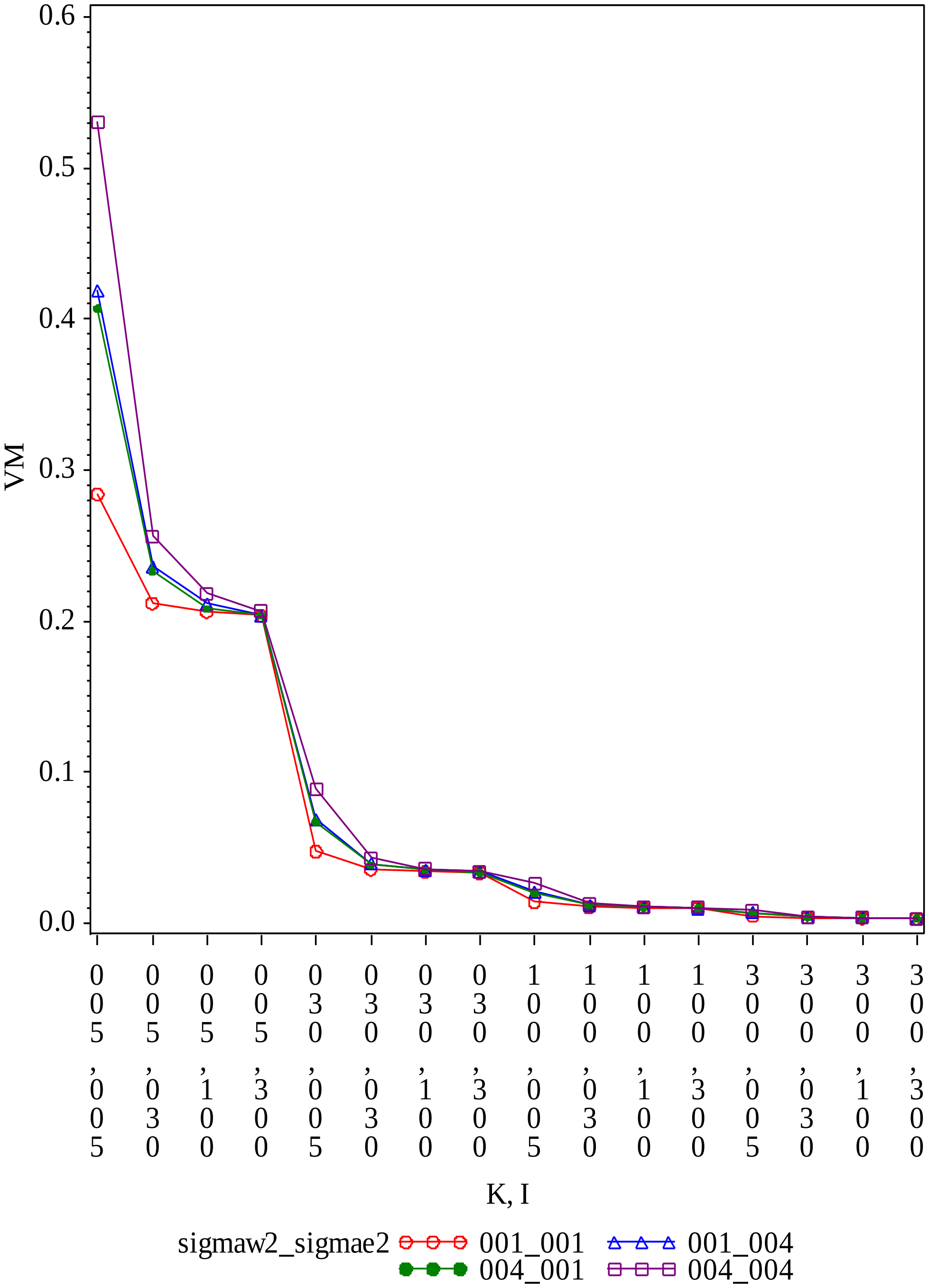

Figure 5.13: V_M vs. sample size K and I by σ_w^2 and σ_e^2 (where σ_b^2 =1, σ_w^2 = 1 or 4, and σ_e^2 = 1 or 4)

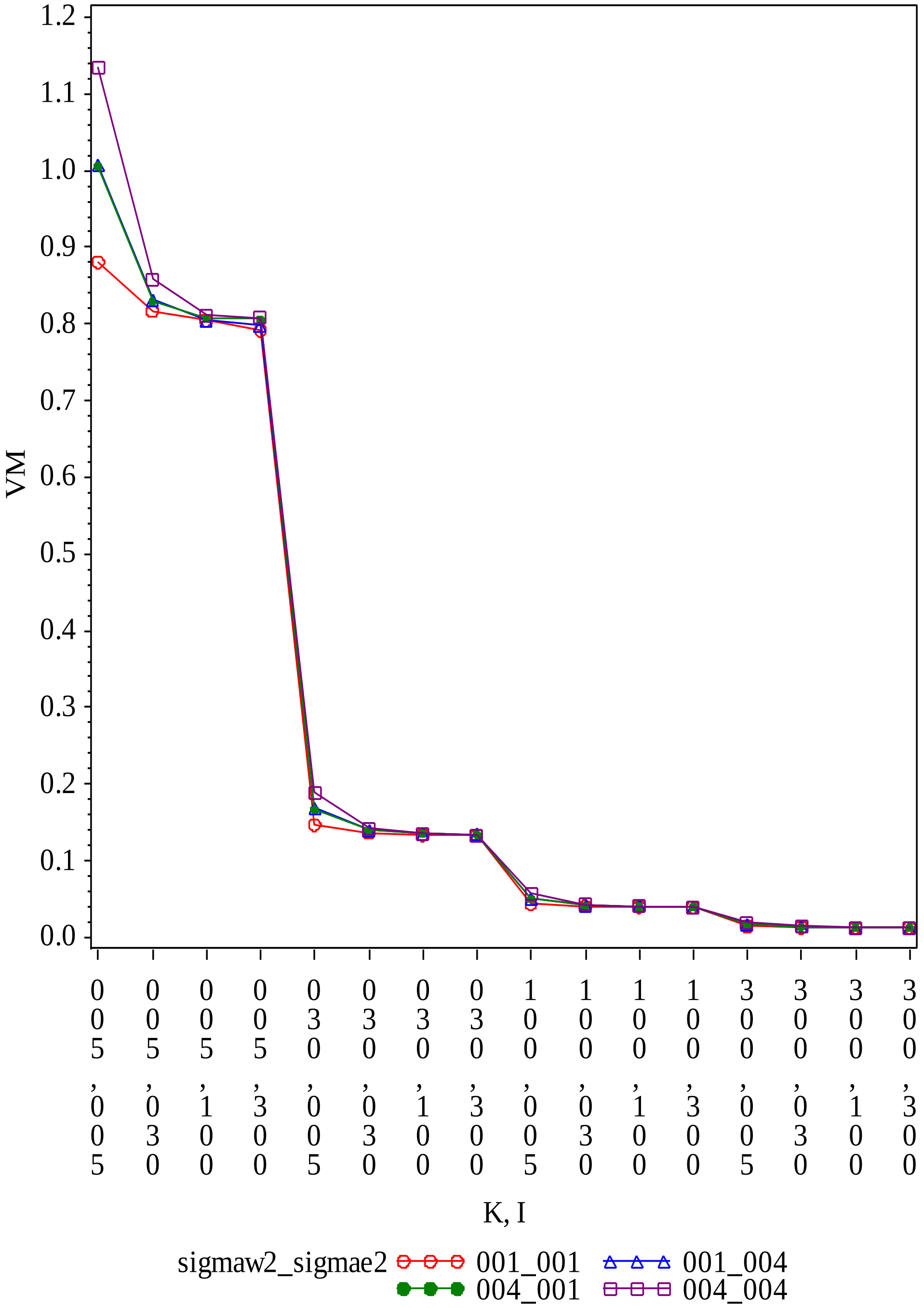

Figure 5.14: V_M vs. sample size *K* and *I* by σ_w^2 and σ_e^2 (where σ_b^2 =4, σ_w^2 = 1 or 4, and σ_e^2 = 1 or 4)

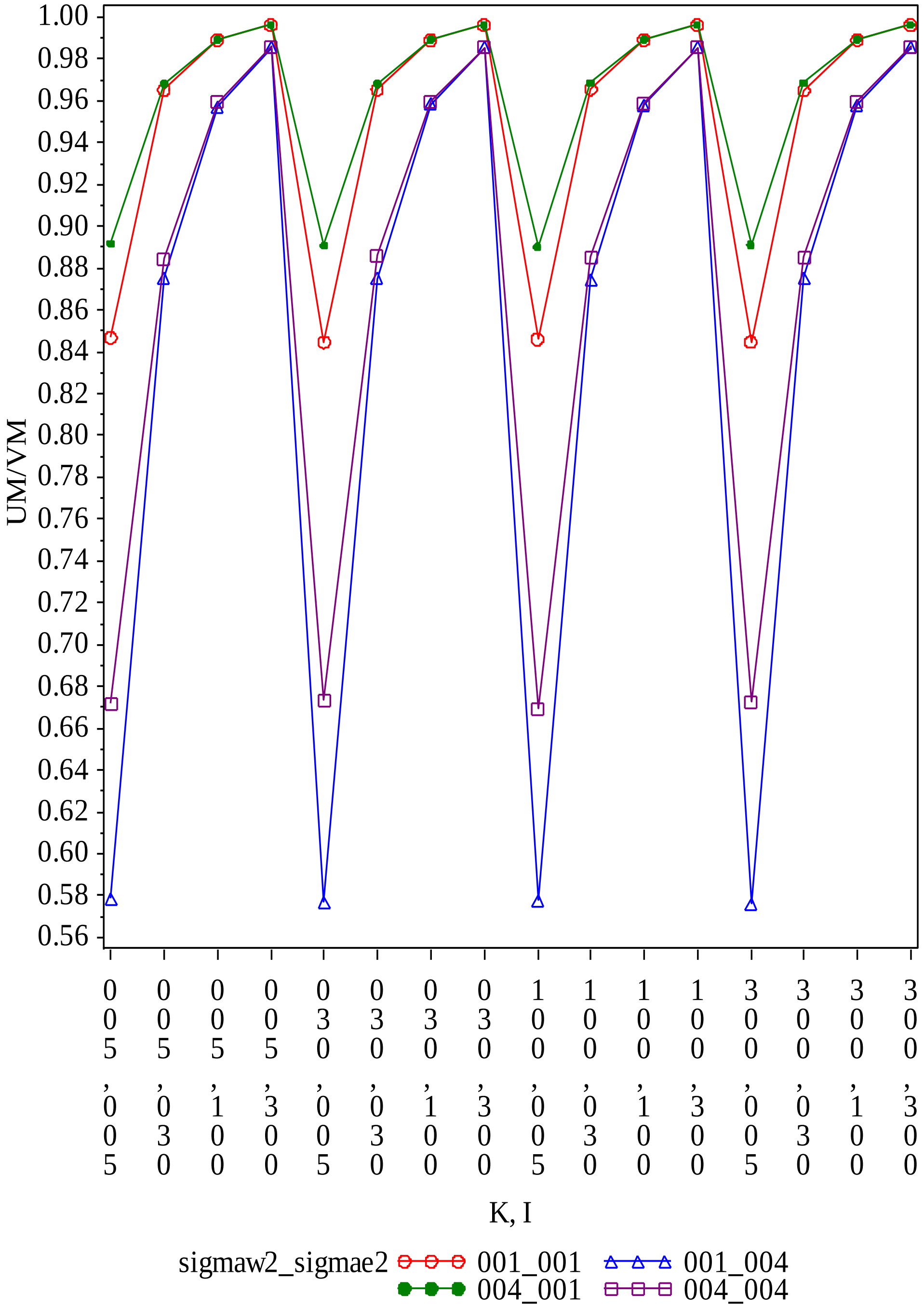

Figure 5.15: U_M / V_M vs. sample size K and I by σ_w^2 and σ_e^2 (where σ_b^2 =1, σ_w^2 = 1 or 4, and σ_e^2 = 1 or 4)

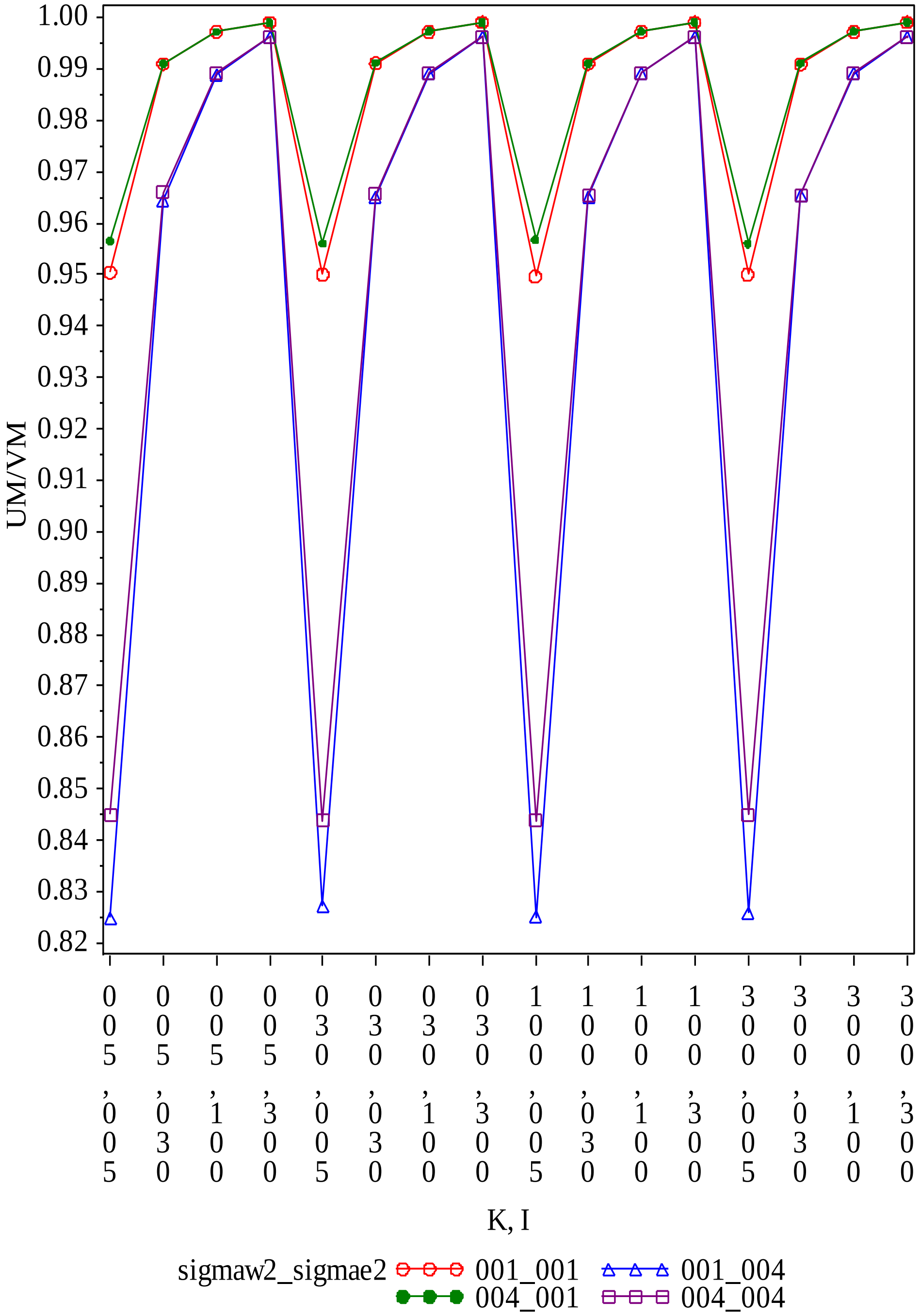

Figure 5.16: U_M / V_M vs. sample size K and I by σ_w^2 and σ_e^2 (where $\sigma_b^2 = 4$, $\sigma_w^2 = 1$ or 4, and $\sigma_e^2 = 1$ or 4)

Regression analyses on sampling variances

Residual analysis of the regression models indicated the existence of outliers, non-normal residual distribution and non-constant residual variance. Table 5.17

summarized the distribution of the residuals. The inference from the regression model is based on the fact that the F statistic is robust against deviation from normal distribution and homogeneity of variances.

Table 5.17 Outlying residuals, skewness and kurtosis in the regression analysis, when the variance components take values 1 and 4

Outcome	Predictor						Distribution statistics	
	σ_b^2	σ_w^2	σ_e^2	K	I	Residual outliers	Skewness	Kurtosis
V_M	4	4	4	5	5	0.169	1.86	6.40
	1	4	4	5	5	0.163		
	1	1	4	5	5	0.097		
	4	1	4	5	5	0.090		
	4	4	1	5	5	0.089		
	1	4	1	5	5	0.089		
U_M	4	4	1	5	5	0.083	2.21	8.10
	1	4	1	5	5	0.082		
	4	4	4	5	5	0.081		
	1	4	4	5	5	0.076		
B_M	1	1	4	5	5	0.080	2.22	8.07
	4	1	4	5	5	0.080		
	4	4	4	5	5	0.080		
	1	4	4	5	5	0.079		
U_M/V_M	1	1	4	5	5	-0.069	-0.77	1.53
	1	1	4	100	5	-0.069		
	1	1	4	30	5	-0.070		
	1	1	4	300	5	-0.071		

When the variance components takes the value of 4, rather than 100, most relationships between the outcome variables and the simulation factors still holds in terms of whether they are significant or not. A few parameter estimates changed from significant to non-significant and are shown in italic font and underlined in the table. Specifically, they are the predictors for V_M: $\sigma_w^2 * K$, $\sigma_w^2 * K^2$, $\sigma_w^2 * I$, $\sigma_w^2 * I^2$, $\sigma_e^2 * K^2$, and $\sigma_e^2 * I^2$.

By examining semipartial $\hat{\eta}^2$, we can identify predictors having major impact on the outcome variables, after partialling out other predictors. For V_M and U_M, σ_b^2

has the largest impact (about 15% of the total variance), and along with its interactions with *K*, over 34% of the variance explained by the model (R^2=0.9834 or 0.9948) or the total variance is accounted for. For B_M, σ_e^2 has the largest impact (9.8% of the total variance), and along with its interactions with *K* and *I*, over 46% of the variance explained by the model (R^2=0.5664) or around 26% of the total variance is accounted for.

Table 5.18 Parameter estimates, *P*-values of *F*-tests and semipartial $\hat{\eta}^2$'s for the regression models on U_M, B_M and V_M, when the ratio of the variance components is 4

Outcome Variables	V_M (R^2=0.9834)			U_M (R^2=0.9948)			B_M (R^2=0.5664)		
Parameter	Estimate	*P*-value	$\hat{\eta}^2$	Estimate	*P*-value	$\hat{\eta}^2$	Estimate	*P*-value	$\hat{\eta}^2$
σ_b^2	**0.1494**	**<0.001**	**0.148**	**0.1493**	**<0.001**	**0.164**	1.58E-05	0.995	<0.001
σ_w^2	**0.0115**	**0.014**	**0.001**	**0.0116**	**<0.001**	**0.001**	-2.59E-05	0.991	<0.001
σ_e^2	**0.0125**	**0.008**	**0.001**	-2.01E-04	0.935	<0.001	**0.0115**	**<0.001**	**0.098**
K	**-0.0114**	**<0.001**	**0.031**	**-0.0101**	**<0.001**	**0.027**	**-0.0012**	**0.008**	**0.032**
K^2	**1.23E-04**	**<0.001**	**0.022**	**1.07E-04**	**<0.001**	**0.019**	**1.38E-05**	**0.014**	**0.027**
K^3	**-2.95E-07**	**<0.001**	**0.020**	**-2.57E-07**	**<0.001**	**0.017**	**-3.42E-08**	**0.017**	**0.025**
I	**-0.0025**	**<0.001**	**0.003**	**-0.0012**	**0.001**	**0.001**	**-0.0012**	**<0.001**	**0.056**
I^2	**2.88E-05**	**<0.001**	**0.002**	**1.36E-05**	**0.001**	**0.001**	**1.38E-05**	**0.001**	**0.053**
I^3	**-7.15E-08**	**<0.001**	**0.002**	**-3.39E-08**	**0.001**	**0.001**	**-3.42E-08**	**0.001**	**0.051**
$K*I$	**9.48E-07**	**0.001**	**0.002**	**4.50E-07**	**0.002**	**0.001**	**4.52E-07**	**0.002**	**0.045**
$\sigma_b^2 * K$	**-0.0066**	**<0.001**	**0.088**	**-0.0066**	**<0.001**	**0.098**	-6.36E-07	0.997	<0.001
$\sigma_b^2 * K^2$	**6.96E-05**	**<0.001**	**0.058**	**6.96E-05**	**<0.001**	**0.064**	7.02E-09	0.997	<0.001
$\sigma_b^2 * K^3$	**-1.66E-07**	**<0.001**	**0.050**	**-1.66E-07**	**<0.001**	**0.056**	-1.70E-11	0.997	<0.001
$\sigma_b^2 * I$	-1.42E-06	0.985	<0.001	-1.04E-06	0.979	<0.001	-3.47E-07	0.993	<0.001
$\sigma_b^2 * I^2$	3.79E-09	0.987	<0.001	2.70E-09	0.983	<0.001	9.89E-10	0.994	<0.001
$\sigma_w^2 * K$	*-1.13E-04*	*0.126*	*<0.001*	**-1.14E-04**	**0.004**	**<0.001**	3.88E-07	0.992	<0.001
$\sigma_w^2 * K^2$	*3.03E-07*	*0.198*	*<0.001*	**3.04E-07**	**0.016**	**<0.001**	-1.08E-09	0.993	<0.001
$\sigma_w^2 * I$	*-1.13E-04*	*0.126*	*<0.001*	**-1.14E-04**	**0.004**	**<0.001**	2.71E-07	0.994	<0.001
$\sigma_w^2 * I^2$	*3.07E-07*	*0.191*	*<0.001*	**3.08E-07**	**0.015**	**<0.001**	-7.07E-10	0.995	<0.001
$\sigma_e^2 * K$	*-1.18E-04*	*0.110*	*<0.001*	2.83E-06	0.942	<0.001	**-1.10E-04**	**0.005**	**0.035**
$\sigma_e^2 * K^2$	*3.15E-07*	*0.179*	*<0.001*	-7.63E-09	0.951	<0.001	**2.94E-07**	**0.018**	**0.025**
$\sigma_e^2 * I$	*-1.21E-04*	*0.104*	*<0.001*	4.32E-07	0.991	<0.001	**-1.10E-04**	**0.005**	**0.035**
$\sigma_e^2 * I^2$	*3.23E-07*	*0.169*	*<0.001*	2.94E-10	0.998	<0.001	**2.93E-07**	**0.018**	**0.025**
$\sigma_b^2 * \sigma_w^2$	6.73E-05	0.940	<0.001	6.45E-05	0.891	<0.001	2.48E-06	0.996	<0.001
$\sigma_b^2 * \sigma_e^2$	-3.41E-05	0.969	<0.001	-3.71E-05	0.937	<0.001	2.74E-06	0.995	<0.001
$\sigma_w^2 * \sigma_e^2$	1.89E-05	0.983	<0.001	2.50E-05	0.958	<0.001	-5.51E-06	0.991	<0.001

Regression analyses were carried out for the ratio variable of U_M over V_M. None of the predictors involving *K* are significant, while all other predicators are significant except $\sigma_b^2 * \sigma_w^2$ and $\sigma_w^2 * \sigma_e^2$. The difference to the cases with the extreme ratio of the variance components is that the variance components play a less important role in explaining the ratio variable and the cluster size becomes more important - σ_b^2 $*\sigma_w^2$ and $\sigma_w^2 * \sigma_e^2$ become non-significant, $\sigma_b^2 * \sigma_e^2$ explains less variation of the ratio, $\sigma_b^2 * I^2$ become significant, and the terms *I, I^2* and I^3 become major predictors.

Semipartial $\hat{\eta}^2$ in the table shows that the major impact to the ratio is from σ_e^2 (0.134), *I* (0.168), I^2 (0.121), and I^3(0.104),.

Table 5.19 Parameter estimates, *P*-values of *F*-tests and semipartial $\hat{\eta}^2$'s for the regression model on the ratio variables U_M / V_M, when the ratio of the variance components is 4

Outcome Variables	U_M/V_M (R^2=0.9433)		
Parameter	Estimates	*P*-value	$\hat{\eta}^2$
σ_b^2	**0.0266**	**0.000**	**0.039**
σ_w^2	**0.0086**	**0.004**	**0.005**
σ_e^2	**-0.0454**	**0.000**	**0.134**
K	-3.95E-07	0.999	0.000
K^2	-4.91E-09	0.999	0.000
K^3	6.99E-12	1.000	0.000
I	**0.0069**	**0.000**	**0.168**
I^2	**-7.00E-05**	**0.000**	**0.121**
I^3	**1.65E-07**	**0.000**	**0.104**
$K*I$	2.18E-09	0.990	0.000
$\sigma_b^2 * K$	1.91E-06	0.992	0.000
$\sigma_b^2 * K^2$	-2.11E-08	0.993	0.000
$\sigma_b^2 * K^3$	5.15E-11	0.993	0.000
$\sigma_b^2 * I$	**-4.11E-04**	**0.000**	**0.043**
$\sigma_b^2 * I^2$	***1.06E-06***	***0.000***	***0.028***
$\sigma_w^2 * K$	-8.40E-07	0.986	0.000
$\sigma_w^2 * K^2$	3.41E-09	0.982	0.000
$\sigma_w^2 * I$	**-1.27E-04**	**0.008**	**0.004**
$\sigma_w^2 * I^2$	**3.43E-07**	**0.023**	**0.003**
$\sigma_e^2 * K$	-5.55E-07	0.991	0.000
$\sigma_e^2 * K^2$	2.21E-09	0.988	0.000
$\sigma_e^2 * I$	**5.13E-04**	**0.000**	**0.067**
$\sigma_e^2 * I^2$	**-1.33E-06**	**0.000**	**0.045**
$\sigma_b^2 * \sigma_w^2$	*-0.0010*	*0.078*	*0.002*
$\sigma_b^2 * \sigma_e^2$	**0.0034**	**0.000**	**0.020**
$\sigma_w^2 * \sigma_e^2$	*5.45E-04*	*0.338*	*0.001*

5.5.3 Research Question 3: What are the relationships between the sampling variance of the imputations and those of the true score and the observed score?

Using the same set of 5,000 repeated samples just used to examine research question 2, we compared the sampling variance of mean estimates based on the imputed score and those of the true score and the observed score.

5.5.3.1 The ratio of the variance components is 100

Table 5.20 shows the ratio of the sampling variance based on the observed score over that of the true score, when the variance components take values 1 and 100. As shown analytically in formula (5.13) in section 5.4.2, the sampling variance based on the observed score is expected to be larger than that of the true score. For the 128 combinations in the table, the ratios are larger than one, with only one exception, as shown in italic font and underlined in the table. The exceptional case is due to the sampling error of the estimate. The ratio is very close to 1 when the error variance is much smaller than other variance components.

Table 5.20 The ratio of the sampling variance based on the observed score over that of the true score, when the variance components take values 1 and 100

K	*I*	1/1/1	1/1/100	1/100/1	1/100/100	100/1/1	100/1/100	100/100/1	100/100/100
5	5	1.1709	17.7765	1.0111	1.9453	1.0019	1.2040	1.0027	1.1602
5	30	1.0330	4.1887	1.0087	1.7522	1.0008	1.0319	1.0008	1.0298
5	100	1.0097	1.9791	1.0035	1.4853	1.0001	1.0094	*0.999998*	1.0087
5	300	1.0036	1.3308	1.0025	1.2604	1.0001	1.0020	1.0002	1.0037
30	5	1.1677	17.6783	1.0095	1.9561	1.0017	1.2002	1.0017	1.1701
30	30	1.0320	4.2128	1.0076	1.7617	1.0004	1.0335	1.0003	1.0319
30	100	1.0100	1.9940	1.0049	1.5109	1.0002	1.0102	1.0002	1.0106
30	300	1.0036	1.3330	1.0024	1.2473	1.0000	1.0036	1.00003	1.0032
100	5	1.1647	17.6996	1.0102	1.9533	1.0020	1.1982	1.0015	1.1663
100	30	1.0327	4.2439	1.0079	1.7743	1.0004	1.0330	1.0003	1.0317
100	100	1.0102	1.9933	1.0048	1.5039	1.0001	1.0095	1.0001	1.0099
100	300	1.0033	1.3295	1.0026	1.2507	1.0000	1.0034	1.00004	1.0033
300	5	1.1664	17.6748	1.0094	1.9526	1.0020	1.1989	1.0016	1.1663
300	30	1.0318	4.2274	1.0078	1.7706	1.0003	1.0331	1.0003	1.0324
300	100	1.0100	1.9856	1.0051	1.5007	1.0001	1.0099	1.0001	1.0100
300	300	1.0032	1.3326	1.0025	1.2500	1.00003	1.0033	1.00003	1.0033

Table 5.21 shows the ratio of the sampling variance based on the imputed data V_M over that of the observed score, when the variance components take values 1 and 100. For the 128 combinations, the ratios are larger than one, with eight exceptions. This result suggests that the sampling variance based on the imputed score is expected to be larger than the observed score. The exceptional cases are suspected to be due to the sampling error of the estimate. In the following paragraphs, we will explore this point by examining the trend of the number of exceptional cases after changing the number of repeated samples. The ratio is very close to 1 when the error variance is much smaller than other variance components.

Table 5.21 The ratio of the sampling variance based on the imputed data over that of the observed score, when the variance components take values 1 and 100

K	*I*	1/1/1	1/1/100	1/100/1	1/100/100	100/1/1	100/1/100	100/100/1	100/100/100
5	5	1.0155	1.0986	1.0007	1.0528	1.00005	1.0179	*0.9998*	1.0173
5	30	1.0019	1.0852	1.0007	1.0418	1.00003	1.0035	1.00001	1.0034
5	100	1.0012	1.0540	1.0003	1.0396	1.00002	1.0005	*0.99999*	1.0008
5	300	*0.9998*	1.0296	1.0001	1.0183	*0.99999*	*0.9997*	*0.99998*	1.0002
30	5	1.0133	1.0943	1.0006	1.0420	1.0003	1.0176	1.0002	1.0142
30	30	1.0030	1.0734	1.0006	1.0406	1.0001	1.0033	1.0001	1.0041
30	100	1.0008	1.0502	1.0007	1.0331	1.00002	1.0011	*0.99998*	1.0007
30	300	1.0004	1.0248	1.0003	1.0207	1.00001	1.0004	1.00001	1.0003
100	5	1.0135	1.0913	1.0010	1.0496	1.0002	1.0181	1.0002	1.0151
100	30	1.0030	1.0788	1.0008	1.0489	1.00003	1.0033	1.00004	1.0030
100	100	1.0010	1.0504	1.0004	1.0350	1.00001	1.0010	1.00001	1.0009
100	300	1.0003	1.0264	1.0003	1.0202	*0.999998*	1.0003	1.000002	1.0003
300	5	1.0132	1.0898	1.0009	1.0531	1.0002	1.0159	1.0002	1.0145
300	30	1.0028	1.0811	1.0008	1.0412	1.00003	1.0033	1.00003	1.0032
300	100	1.0010	1.0469	1.0005	1.0345	1.00001	1.0009	1.00001	1.0010
300	300	1.0003	1.0249	1.0002	1.0214	1.000002	1.0004	1.00001	1.0003

To show that the ratio statistics in table 5.20 and 5.21 converge to the observed pattern in terms of the number of repeated samples, the simulation study was carried out using 1,000, 5000, and 25,000 repeated samples. The tables 5.22 and 5.23 list simulation conditions for the exceptional cases for the 5 sets of simulations with 1,000 repeated samples, for the 5 sets of simulations with 5,000 repeated samples, and for the 25,000 repeated samples, which are the combination of the 5 sets of 5,000 repeated samples. The counts of the exceptional cases are shown at the bottom of the tables. As shown in table 5.22, for the ratio of the sampling variance based on the observed score over that of the true score, when there are 1000 repeated samples, the number of exceptional cases ranges between 3 and 7 among the 5 sets of repeated samples; when there are 5,000 repeated samples, the number ranges between 1 and 3 among the 5 sets of repeated samples; and when there are 25,000 repeated samples, there is 0 exceptional case. In table 5.23, for the ratio of the sampling variance based on the imputed data over that of the observed score, when there are 1,000 repeated samples, the number of exceptional cases ranges between 9 and 15 among the 5 sets of repeated samples; when there are 5,000 repeated samples, the number of exceptional cases ranges between 2 and 9 among the 5 sets of repeated samples; and when there are 25,000 repeated samples, there is 1 exceptional case. Note that the 5 sets of 5,000 repeated samples have exceptional cases in different combination of factors and no exceptional case happens for all five sets of repeated samples. The reduction of the number of the exceptional cases and the lack of pattern of these cases suggest that the observed pattern converges in terms of the number of repeated samples, and hence the exceptional cases are due to the sampling error. Further, the exceptional cases are concentrated in the conditions with lower error variance and/or larger cluster size.

Table 5.22 The simulation conditions with the ratio of the sampling variance based on the observed score over that of the true score lower than 1 for the data with 1,000 repeated samples, 5,000 repeated samples and 25,000 repeated samples, when the variance components take values 1 and 100.

σ_b^2	σ_w^2	σ_e^2	K	I	5 sets of 1,000 repeated samples					25,000 Repeated samples (5 sets of 5,000 and combined)					
					1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	Combined
100	1	1	5	5	X				X						
100	100	1	5	5								X			
100	1	1	5	30		X			X			X			
100	100	1	5	30		X		X		X			X		
100	1	1	5	100	X		X		X		X			X	
100	100	1	5	100			X	X	X	X					
100	1	1	5	300	X	X			X					X	
100	100	1	5	300						X					
100	1	1	30	100		X									
100	1	1	30	300					X			X			
100	100	1	30	300		X	X		X						
100	1	1	100	100			X								
100	1	1	100	300	X			X							
100	100	1	100	300		X									
Count					4	6	4	3	7	3	1	3	1	2	0

Table 5.23 The simulation conditions with the ratio of the sampling variance based on the imputation over that of observed scores lower than 1 for the data with 1,000, 5,000 and 25,000 repeated samples, when the variance components take values 1 and 100.

σ_b^2	σ_w^2	σ_e^2	K	I	5 sets of 1,000 repeated samples					25,000 Repeated samples (5 sets of 5,000 and combined)					
					1	2	3	4	5	1	2	3	4	5	Combined
1	100	1	5	5				X							
100	1	1	5	5	X	X	X	X	X	X				X	
100	100	1	5	5								X	X		
100	100	100	5	5				X							
1	1	1	5	30				X							
1	100	1	5	30			X								
100	1	1	5	30	X	X		X		X				X	
100	100	1	5	30	X			X							
100	100	100	5	30		X									
1	1	1	5	100				X							
1	100	1	5	100				X							
100	1	1	5	100	X		X		X					X	
100	1	100	5	100		X	X		X		X				
100	100	1	5	100			X		X		X		X	X	
1	1	1	5	300			X		X						
1	100	1	5	300				X		X			X		
100	1	1	5	300	X	X	X	X		X	X			X	
100	1	100	5	300		X									
100	100	1	5	300	X	X				X		X	X	X	X
100	100	100	5	300		X	X		X						
100	1	1	30	5				X	X						
100	100	1	30	5		X	X	X						X	
100	1	1	30	30									X		
100	1	1	30	100			X	X			X		X		
100	100	1	30	100									X		
1	1	1	30	300	X										
1	100	1	30	300					X						
100	1	1	30	300			X							X	
100	100	1	30	300				X		X			X		
100	1	1	100	30						X					
100	1	1	100	100	X					X	X				
100	100	1	100	100		X		X	X	X					
100	1	1	100	300		X	X							X	
100	100	1	100	300							X		X		
100	100	1	300	100					X						
100	1	1	300	300	X		X								
100	100	1	300	300		X									
					9	12	13	15	10	9	6	2	9	9	1

Next, the number of imputations in MI needed for precise estimation of sampling variance was explored by increasing the number of imputations. As we have just seen, when the number of imputations is 10 and the number of repeated samples is 1000, we observed 9 to 15 exceptional cases. When the number of imputations was increased to 100, the number of exceptional cases was reduced to 5. This result provides evidence for the following point: by increasing the number of imputations, the precision of estimation can be improved, and hence the number of the exceptional cases can be reduced.

To further examine the characteristics of the sampling variance of the mean based on the imputed scores, we created the ratio of the sampling variance of the mean based on the true score to that based on the imputed score V_M, which represents the proportion of $V_M$that is from the true score variance. As it was shown analytically in formula (5.14) and (5.15) in section 5.4.2 that $Var(\bar{\theta})$ and U_M have the same expected value, we expect the analysis results for $Var(\bar{\theta})/V_M$ is the same as U_M/V_M

This ratio variable $Var(\bar{\theta})/V_M$ was plotted against sample sizes *K* and *I*, by the variance components σ_b^2, σ_w^2 and σ_e^2. As expected, the shape of the graphs is almost identical to graphs 5.7 and 5.8 for the ratio U_M/V_M.

Regression analyses were carried out for the ratio variable and the results are shown in table 5.24. The same set of independent variables were used as the analysis for the ratio U_M/V_M. The results for the two ratio variables are almost identical. All three main effects of the variance component variables and their interactions with each other are significant. Details can be seen in the discussion for the ratio U_M/V_M.

Table 5.24 Parameter estimates, *P*-values of *F*-tests and semipartial $\hat{\eta}^2$'s for the ratio variables $Var(\bar{\theta})/V_M$, when the ratio of the variance components is 100

Outcome Variables	$Var(\bar{\theta})/V_M$ ((R^2=0.9580)		
Parameter	Estimate	*P*-value	$\hat{\eta}^2$
σ_b^2	**0.00184**	**<0.001**	**0.018**
σ_w^2	**0.00159**	**<0.001**	**0.016**
σ_e^2	**-0.00656**	**<0.001**	**0.273**
K	4.18E-05	0.971	<0.001
K^2	-4.56E-07	0.975	<0.001
K^3	1.10E-09	0.976	<0.001
I	**0.00405**	**<0.001**	**0.009**
I^2	**-3.36E-05**	**0.002**	**0.004**
I^3	**7.58E-08**	**0.005**	**0.003**
$K*I$	2.13E-09	0.995	<0.001
$\sigma_b^2 * K$	-5.70E-07	0.971	<0.001
$\sigma_b^2 * K^2$	8.38E-09	0.968	<0.001
$\sigma_b^2 * K^3$	-2.20E-11	0.967	<0.001
$\sigma_b^2 * I$	**-1.31E-05**	**0.002**	**0.004**
$\sigma_b^2 * I^2$	2.09E-08	0.111	0.001
$\sigma_w^2 * K$	-1.80E-07	0.965	<0.001
$\sigma_w^2 * K^2$	4.72E-10	0.971	<0.001
$\sigma_w^2 * I$	**-1.40E-05**	**0.001**	**0.005**
$\sigma_w^2 * I^2$	**3.30E-08**	**0.013**	**0.003**
$\sigma_e^2 * K$	-1.85E-07	0.964	<0.001
$\sigma_e^2 * K^2$	5.48E-10	0.966	<0.001
$\sigma_e^2 * I$	**2.35E-05**	**<0.001**	**0.014**
$\sigma_e^2 * I^2$	**-4.85E-08**	**<0.001**	**0.006**
$\sigma_b^2 * \sigma_w^2$	**-1.41E-05**	**<0.001**	**0.021**
$\sigma_b^2 * \sigma_e^2$	**4.25E-05**	**<0.001**	**0.192**
$\sigma_w^2 * \sigma_e^2$	**1.03E-05**	**<0.001**	**0.011**

5.5.3.2 The ratio of the variance components is 4

This section presents the analysis results when the ratio of the variance components is 4. Tables 5.25 presents the ratio of the sampling variance based on the observed score over that of the true score and table 5.26 presents the ratio of the sampling variance based on the imputed data over that of the observed score. With fewer exceptional cases, the tables show the same pattern as observed when ratio of the variance components is 100: the sampling variance based on the imputed score is larger than that of the observed score, which is larger than that of the true score. There is no exceptional case in table 5.25 and 1 case in table 5.26.

Table 5.25 The ratio of the sampling variance based on the observed score over that of the true score, when the variance components take values 1 and 4

K	*I*	1/1/1	1/1/4	1/4/1	1/4/4	4/1/1	4/1/4	4/4/1	4/4/4
5	5	1.16414	1.68143	1.11429	1.43831	1.04286	1.18798	1.04165	1.15848
5	30	1.03107	1.13092	1.02882	1.11378	1.00863	1.03677	1.00723	1.03522
5	100	1.01230	1.03915	1.00996	1.04053	1.00253	1.00972	1.00339	1.00869
5	300	1.00296	1.01190	1.00326	1.01479	1.00055	1.00483	1.00085	1.00500
30	5	1.16462	1.67529	1.11060	1.44657	1.04814	1.19353	1.04050	1.16779
30	30	1.03218	1.13346	1.02941	1.12091	1.00907	1.03198	1.00839	1.03168
30	100	1.00998	1.03956	1.01000	1.03865	1.00269	1.00980	1.00232	1.01048
30	300	1.00310	1.01400	1.00358	1.01267	1.00113	1.00384	1.00084	1.00406
100	5	1.16445	1.66617	1.10987	1.44539	1.04665	1.18993	1.04259	1.16701
100	30	1.03227	1.12835	1.02969	1.11809	1.00818	1.03388	1.00764	1.03268
100	100	1.00975	1.03970	1.00948	1.03803	1.00281	1.00991	1.00253	1.01002
100	300	1.00344	1.01348	1.00322	1.01264	1.00072	1.00341	1.00081	1.00340
300	5	1.16660	1.66174	1.11162	1.44535	1.04758	1.19145	1.04148	1.16636
300	30	1.03199	1.12941	1.02946	1.11696	1.00827	1.03287	1.00795	1.03236
300	100	1.00969	1.03938	1.00986	1.03849	1.00270	1.01035	1.00230	1.01002
300	300	1.00341	1.01312	1.00331	1.01299	1.00086	1.00317	1.00092	1.00353

Table 5.26 The ratio of the sampling variance based on the imputed data over that of the observed score, when the variance components take values 1 and 4

K	*I*	1/1/1	1/1/4	1/4/1	1/4/4	4/1/1	4/1/4	4/4/1	4/4/4
5	5	1.01244	1.02965	1.00885	1.03332	1.00421	1.01581	1.00408	1.01463
5	30	1.00263	1.01377	1.00390	1.01013	1.00074	1.00241	1.00089	1.00200
5	100	1.00029	1.00428	1.00080	1.00332	1.00064	1.00137	1.00002	1.00107
5	300	1.00082	1.00143	1.00022	1.00125	1.00009	1.00054	1.00020	1.00054
30	5	1.01402	1.03807	1.01069	1.02632	1.00447	1.01405	1.00401	1.01704
30	30	1.00331	1.01231	1.00401	1.00952	1.00049	1.00306	1.00084	1.00319
30	100	1.00111	1.00281	1.00072	1.00381	1.00015	1.00131	1.00015	1.00117
30	300	1.00043	1.00121	1.00037	1.00175	1.00006	1.00034	*0.99998*	1.00048
100	5	1.01329	1.04030	1.01144	1.03421	1.00502	1.01843	1.00325	1.01581
100	30	1.00310	1.01198	1.00304	1.01057	1.00087	1.00283	1.00084	1.00328
100	100	1.00076	1.00338	1.00110	1.00378	1.00024	1.00075	1.00024	1.00116
100	300	1.00037	1.00138	1.00015	1.00128	1.00008	1.00036	1.00007	1.00029
300	5	1.01469	1.04119	1.00986	1.02922	1.00466	1.01691	1.00415	1.01434
300	30	1.00366	1.01109	1.00275	1.01050	1.00088	1.00272	1.00090	1.00330
300	100	1.00101	1.00372	1.00084	1.00350	1.00028	1.00108	1.00021	1.00094
300	300	1.00028	1.00127	1.00026	1.00133	1.00010	1.00035	1.00009	1.00027

Table 5.27 presents the parameter estimates, P-values and semipartial $\hat{\eta}^2$ for the ratio variables $Var(\bar{\theta})/V_M$ when the variance components take values 1 and 4.

Outcome Variables	$Var(\bar{\theta})/V_M$ (R^2=0.9432)		
Parameter	Estimate	P-Value	$\hat{\eta}^2$
σ_b^2	**0.02682**	**<0.001**	**0.040**
σ_w^2	**0.00851**	**0.005**	**0.005**
σ_e^2	**-0.04558**	**<0.001**	**0.136**
K	-2.14E-05	0.968	<0.001
K^2	3.26E-07	0.961	<0.001
K^3	-8.79E-10	0.959	<0.001
I	**0.00688**	**<0.001**	**0.167**
I^2	**-6.96E-05**	**<0.001**	**0.120**
I^3	**1.64E-07**	**<0.001**	**0.103**
$K*I$	8.69E-09	0.959	<0.001
$\sigma_b^2 * K$	-1.39E-06	0.994	<0.001
$\sigma_b^2 * K^2$	9.13E-10	1.000	<0.001
$\sigma_b^2 * K^3$	8.27E-12	0.999	<0.001
$\sigma_b^2 * I$	**-0.00041**	**<0.001**	**0.044**
$\sigma_b^2 * I^2$	**1.06E-06**	**<0.001**	**0.028**
$\sigma_w^2 * K$	-1.50E-06	0.975	<0.001
$\sigma_w^2 * K^2$	5.21E-09	0.972	<0.001
$\sigma_w^2 * I$	**-0.00013**	**0.009**	**0.004**
$\sigma_w^2 * I^2$	**3.40E-07**	**0.025**	**0.003**
$\sigma_e^2 * K$	-1.10E-06	0.981	<0.001
$\sigma_e^2 * K^2$	5.29E-09	0.972	<0.001
$\sigma_e^2 * I$	**0.00052**	**<0.001**	**0.068**
$\sigma_e^2 * I^2$	**-1.34E-06**	**<0.001**	**0.045**
$\sigma_b^2 * \sigma_w^2$	-0.00100	0.081	0.002
$\sigma_b^2 * \sigma_e^2$	**0.00339**	**<0.001**	**0.020**
$\sigma_w^2 * \sigma_e^2$	0.00058	0.306	0.001

Chapter 6 : Conclusion

6.1 Importance of the Study

Rubin's MI methodology for handling latent variables in analyzing survey data is widely used in large-scale educational assessments, such as NAEP. This research fills in an important gap in the backing for these procedures, namely the demonstration of properties for imputed latent variables in a random-effects model for two-stage cluster sample designs. Random effects are characteristics of common evaluation scenarios where there is multistage sampling. Large-scale assessments including NAEP use the fixed-effects model in developing plausible values for complex samples. This study provides a framework for including random-effects in the production of plausible values.

The analytic portion of the research provides derivations of expectations of key population parameters in the simple case of known population parameters. The empirical portion extends the construction of imputations to the case of unknown means, and examines the performance of the resulting imputed data sets with simulations. This work provides the two-stage sampling case with the backing that was provided for the fixed-effects covariates case given in Mislevy (1991).

6.2 Major findings

In the case of known population parameters, the imputation based on observations with measurement errors was constructed under the MI framework for a two-stage cluster sample design so as to re-express key characteristics of the latent true score variable. The analytical solution shows that the estimator constructed based on the imputation reproduces these population parameters. This latent variable is assumed to be normally distributed at each of the two levels in the population model, the cluster level and the individual level, and the measurement error is normally distributed. The known parameters include the population mean, cluster means, the within-cluster variance, the between-cluster variance and the variance of the measurement error.

In the case of unknown population and cluster means, a simulation study was carried out to demonstrate properties of the plausible values constructed under the MI framework for a two-stage cluster sample design. A Bayesian procedure with a noninformative prior on population and cluster means was approximated by estimating these population parameters based on the observed data. The simulation study findings are summarized below:

Research Question 1:

According to the empirical study based on the simulated data, the sample estimator based on imputed scores is unbiased in estimating the population mean, cluster means, total variance, within-cluster variance and between-cluster variance. In contrast, the sample estimator based on the observed score is positively biased in estimating total variance and within-cluster variance, while the bias may be ignorable in rare cases. However, the estimate of the population mean, cluster means, and the variance of cluster means based on the observed score doesn't appear to be biased.

To obtain unbiased point estimates, the variance reconstruction terms were incorporated into the imputed score. The variance of the variance reconstruction terms are defined as $Var(g_k) = (1-\lambda)\sigma_b^2$ and $Var(f_{ik}) = (1-\rho)\sigma_w^2$ and can be shown as

$$Var(g_k) = \frac{\sigma_b^2 \times \frac{(\sigma_w^2 + \sigma_e^2)}{I}}{\left[\sigma_b^2 + \frac{(\sigma_w^2 + \sigma_e^2)}{I}\right]}$$

and

$$Var(f_{ik}) = \frac{\sigma_w^2 \times \sigma_e^2}{(\sigma_w^2 + \sigma_e^2)}.$$

Larger values of the posterior variance at the cluster level $Var(g_k)$ correspond to larger values of any of the three variance components, between-cluster variance (σ_b^2), within-cluster cluster variance (σ_w^2) and measurement error variance (σ_e^2). Larger values of σ_w^2 and σ_e^2 correspond to smaller values of λ and larger proportions of σ_b^2 added from the random component g_k to the variance of the imputed score. The proportion of this added variance over the variance of the imputed score has a complex relationship to these variance components. Nevertheless, $Var(g_k)$ accounts for more than 5% of the overall variance only in a few cases, either when (σ_b^2 , σ_w^2, σ_e^2) = (1,1,100), when (σ_b^2 , σ_w^2, σ_e^2) = (1,1,4) and the cluster size is 30, or when the cluster size is 5.

In constructing individual scores, a larger values of σ_e^2 corresponds to a smaller value of ρ and a larger proportion of σ_w^2 added from the random component f_{ik} to the variance of the imputed score. A larger variance of f_{ik} is associated with larger values of σ_e^2 and σ_w^2, while a larger relative variance $R\_Var(f_{ik})$ is associated with larger values of σ_e^2 and smaller values of σ_b^2 . $Var(f_{ik})$ accounts for a large percentage of the overall variance for the following cases: 25% when $\sigma_b^2 = \sigma_w^2 = \sigma_e^2$ and 49.5% when σ_b^2 = 1and σ_e^2 = 100. The smallest percentage is 0.5% when σ_b^2 = 100 and σ_e^2 = 1. When the ratio of the variance components is 4, $Var\left(f_{ik}\right)$ accounts for 25% of the overall variance when $\sigma_b^2 = \sigma_w^2 = \sigma_e^2$, 40.0% when σ_b^2 = 1 and σ_e^2 = 4, and 10% as the smallest percentage when σ_b^2 = 4 and σ_e^2 = 1.

A larger cluster size corresponds to a larger λ, hence a smaller variance of the random component g_k and a smaller relative variance. However, the variance of f_{ik} is not affected by sample sizes.

Research Question 2:

The relationship between the sampling variance of the mean of the imputed score and the simulation factors were evaluated through tables, graphs and regression analyses.

Simulation results were examined separately for the two simulated cases where the ratio of the variance components was 100 or 4, and the general patterns appeared to be similar. It was shown that the positive impact on V_M from σ_b^2 and σ_w^2 comes through U_M, while the positive impact from σ_e^2 comes through B_M. The impact from σ_b^2 on V_M is much larger than that of σ_w^2 or σ_e^2.

The relationships between the sample sizes and the sampling variance of the imputed score are in a curved shape, as shown in the graphs. The shapes are illustrated by significant negative main effects, positive quadratic terms and negative cubic terms for the number of clusters (*K*) and the cluster size (*I*) in the regression model on V_M, U_M and B_M. In addition, the interaction term *K*I* shows that the impact from *K* (or *I*) on V_M, U_M and B_M is larger when *I* (or *K*) is higher.

According to the interaction between the variance components and the sample sizes, the impact from the sample sizes on the sampling variance is larger when the variance components are larger. σ_b^2 has no significant interaction effect with *I*.

The variance components have no significant interaction effect with each other.

We have identified predictors with major impact to the sampling variance, after partialling out the effect from other predictors. For outcome variables V_M and U_M, σ_b^2 has the largest impact, explaining over 26% of the total variance of the outcome variable when the ratio of the variance components is 100 and around 15% when the ratio is 4. Along with its interactions with *K*, over 60% of the variance is accounted for when the ratio is 100 and over 34% when the ratio is 4. For B_M, σ_e^2 has the largest impact, explaining over 17% of the total variance of the outcome variable when the ratio is 100 and 9.8% when the ratio is 4. Along with its interactions with *K* and *I*, over 38% of this variance is accounted for when the ratio is 100 and around 26% when the ratio is 4.

The relationship between the ratio of U_M over V_M and the simulation factors were examined. This ratio shows the proportion of the sampling variance accounted for by single imputation rather than multiple imputation and the supplement of this ratio reflects the proportion of the missing information due to not observing the true score directly. The variance components σ_b^2 and σ_w^2 have positive effects on the ratio, while σ_e^2 has a negative effect. The interaction effect of $\sigma_b^2 * \sigma_w^2$ is negative and the interaction effects of $\sigma_b^2 * \sigma_e^2$ and $\sigma_w^2 * \sigma_e^2$ are positive. These interaction effects become weak when the ratio of the variance components is 4, rather than 100.

The effect of the cluster size is a curved shape, described by the positive main effect, the negative quadratic term, and the positive cubic term. The interaction between the cluster size and the variance components shows that the effect from the cluster size is weaker when the variance components are at a higher level (100 or 4). The number of clusters has no impact on the ratio.

According to the proportion of total variance explained by the factor of interest, the major impact on the ratio is from σ_e^2 (27.36%) and $\sigma_b^2 * \sigma_e^2$ (19.19%) when the ratio of the variance components is 100, and is from σ_e^2 (13.4%), *I* (16.8%), I^2 (12.1%), and I^3(10.4%) when the ratio is 4.

Research Question 3:

The sampling variance based on the observed score is expected to be the upper boundary of that based on the true score. The cases that violate this expectation are empirically shown to be due to sampling error – the number of cases in violation decreases when the number of repeated samples increases.

The sampling variance based on the imputed score is expected to be the upper boundary of that based on the observed score. It is also empirically shown that violations to this expectation are due to sampling error –again the number of cases in violation decreases when the number of repeated samples increases. In addition, when the number of imputations used in MI increases, fewer cases violate the expectation. The ratio of the sampling variance of the mean based on the true score to that based on the imputed score V_M represents the proportion of $V_M$that comes from the true score variation. The true score sampling variance and U_M are shown to have the same expected value. The analysis on the ratio has almost identical result as U_M/V_M.

6.3 Limitations and Future Research

The simulation study above assumes unknown population mean, but known variance components, σ_b^2 , σ_w^2 and σ_e^2 . However, in practice, the variance components must be estimated from previous research and current data. The proposed research can be extended to the case where the variance components are unknown, and must be estimated.

This study uses a straightforward measurement model – CTT. Simulation study could be extended to more complex models, such as IRT or latent class model, to provide more direct linkage to large-scale assessments. Using the framework provided in this study, which includes random-effects in the production of plausible values, one can investigate biases that may occur in the fixed-effects PVs for various statistics in secondary analyses, including hierarchical analyses, and determine whether incorporating random effects into production models is merited to mitigate them.

A normal distribution is assumed for each random component in this study, and this assumption does in fact accord with the way the data were generated in the simulations. This assumption can be evaluated in practice and simulation study could be extended to other generating distributions for the error variance and the true score variances at different stages, in order to examine robustness of inference to misspecification of distributions.

A combined model with both stratification (fixed effects) and multiple levels of sampling (random effects) could be developed as well. More ambitiously, one could consider the challenge of creating plausible values for multi-level models with predictors at multiple levels.

Another component of complex sampling, unequal weights, could also be investigated in the sampling model. When sampling weights are relevant, the issue of whether to include them in the estimation of the population model and the measurement model can be investigated.

Another interesting extension would be a study of the analytical solution of the sampling variance for simplified cases and the estimates of empirical sampling variance for more complex cases.

There are close connections between plausible values and the augmented-data draws for latent variables in Markov Chain Monte Carlo (MCMC) Bayesian estimation models. Working out these relationships explicitly would be of interest.

Glossary

Symbol	Label
μ	Population mean
ν_k	Population mean for cluster *k*
θ_{ik}	Individual true score for person *i* in cluster *k*
X_{ik}	Individual observed score for person *i* in cluster *k*
σ_b^2	Population between-cluster variance
σ_w^2	Population within-cluster variance
σ_e^2	Population error variance
K	Number of clusters
I	Cluster size
$\hat{\mu}$	Sample mean
$V_{\mu(x)}$	Variance of the sample mean
$\tilde{\mu}_{(m)}$	Imputed overall mean for the imputed pseudo dataset *m,* a random number drawn from $N\left(\hat{\mu}, V_{\mu(x)}\right)$
$\tilde{\nu}_{k(m)}$	Imputed cluster mean for cluster *k* for the imputed pseudo dataset *m*
$\tilde{\theta}_{ik(m)}$	Imputed individual score for person *i* in cluster *k* for the imputed pseudo dataset *m*
RE	The efficiency when using a finite number of proper imputations, *m*, rather than an infinite number
$g_{k(m)}$	Variance reconstruction term at the cluster mean level, drawn from $N(0,(1-\lambda)\sigma_b^2)$
$f_{ki(m)}$	Variance reconstruction term at the individual level within clusters, drawn from $N(0,(1-\rho)\sigma_w^2)$
$\bar{x}_k$	The sample mean of cluster *k*
ρ	Reliability coefficient at the individual level, $\rho = \sigma_w^2 / (\sigma_w^2 + \sigma_e^2)$
λ	Reliability coefficient at the cluster level, $\lambda = \sigma_b^2 / [\sigma_b^2 + (\sigma_w^2 + \sigma_e^2)/I]$
S	The population characteristics
$s_{(m)}$	The point estimates calculated based on imputation data set *m*
s_M	The point estimates averaged across the all imputed pseudo data sets, $s_M = \frac{\sum s_{(m)}}{M}$
$\bar{s}_M$	Average of s_M across 1000 repeated samples
$Var(s_M)$	Variance of s_M across 1000 repeated samples
$z-value$	The standardized score for the mean of s_M across 1000 repeated samples, $z-value = \frac{\bar{s}_M - S}{\sqrt{Var(s_M)/1000}}$

$\overline{\overline{\tilde{\theta}}}_{(m)}$	Estimator of population mean based on the imputed score ($\tilde{\theta}_{ik(m)}$) in the imputed pseudo dataset *m*, $\overline{\overline{\tilde{\theta}}}_{(m)} = \frac{\sum_{k}\sum_{i}\tilde{\theta}_{ik(m)}}{K \times I}$
$\overline{\overline{\theta}}$	Estimator of population mean based on the true score (θ_{ik}), $\overline{\overline{\theta}} = \frac{\sum_{k}\sum_{i}\theta_{ik}}{K \times I}$
$\overline{\overline{x}}$	Estimator of population mean based on the observed score **(** X_{ik} **)**, $\overline{\overline{x}} = \frac{\sum_{k}\sum_{i}x_{ik}}{K \times I}$
$\overline{\tilde{\theta}}_{k(m)}$	Estimator of cluster means based on the imputed score ($\tilde{\theta}_{ik(m)}$) in the imputed pseudo dataset *m*, $\overline{\tilde{\theta}}_{k(m)} = \frac{\sum_{i}\tilde{\theta}_{ik(m)}}{I}$
$\overline{\theta}_{k}$	Estimator of cluster means based on the true score (θ_{ik}), $\overline{\theta}_{k} = \frac{\sum_{i}\theta_{ik}}{I}$
$\overline{x}_{k}$	Estimator of cluster means based on the observed score **(** X_{ik} **)**, $\overline{x}_{k} = \frac{\sum_{i}x_{ik}}{I}$
$Var(\tilde{\theta}_{ik(m)} \mid k)$	Estimator of within-cluster variance based on the imputed score ($\tilde{\theta}_{ik(m)}$) in the imputed pseudo dataset *m*, $Var(\tilde{\theta}_{ik(m)} \mid k) = \frac{\sum_{k}\sum_{i}(\tilde{\theta}_{ik(m)} - \overline{\tilde{\theta}}_{k(m)})^{2}}{K(I-1)}$
$Var(\theta_{ik} \mid k)$	Estimator of within-cluster variance based on the true score (θ_{ik}), $Var(\theta_{ik} \mid k) = \frac{\sum_{k}\sum_{i}(\theta_{ik} - \overline{\theta}_{k})^{2}}{K(I-1)}$
$Var(x_{ik} \mid k)$	Estimator of within-cluster variance based on the observed score **(** X_{ik} **)**,$Var(x_{ik} \mid k) = \frac{\sum_{k}\sum_{i}(x_{ik} - \overline{x}_{k})^{2}}{K(I-1)}$
$Var(\overline{\tilde{\theta}}_{k(m)})$	Variance of cluster means of the imputed score ($\tilde{\theta}_{ik(m)}$) in the imputed pseudo dataset *m*
$Var(\overline{\theta}_{k})$	Variance of cluster means based on the true score (θ_{ik})
$Var(\overline{x}_{k})$	Variance of cluster means based on the true score (X_{ik})
$Var(\tilde{\theta}_{ik})$	Variance of individual scores based on the imputed score ($\tilde{\theta}_{ik(m)}$) in

	the imputed pseudo dataset *m*
$Var(\theta_{ik})$	Variance of individual scores based on the true score (θ_{ik})
$Var(x_{ik})$	Variance of individual scores based on the true score (X_{ik})
$U_{(m)}$	The sampling variance of the point estimate using the plausible values treating the imputed values as observed, calculated based on imputation data set m
U_M	Within imputation variance of the point estimates using the plausible values from multiple imputations, averaged across the all imputed pseudo data sets, $U_M = \frac{\sum U_{(m)}}{M}$
B_M	Between imputation variance of the point estimates for the plausible values from multiple imputations, $B_M = \frac{\sum (s_{(m)} - s_M)^2}{(M-1)}$
V_M	Total variance of the point estimates for the plausible values from multiple imputations, $V_M = U_M + (1+1/M)B_M$
s_1^2	The estimated between-cluster unit variance
s_2^2	The estimated within-cluster unit variance
Bias	The bias is the difference of the estimate and the true population value, $Bias = s_M - S$

Appendix A.

An example of R code – this set of R code was used to create data and conduct analysis for research question 1. The code for research Question 2 and 3 used similar algorithm, with limited revisions.

```
rm(list=ls())

setwd("C:\\Tiandong\\Dissertation\\Simulation\\R\\08_01_11\\V100_Rep1K")

table.final=c();

j=0;

for (k in c(5,30,100,300))     {

for (i in c(5,30,100,300))     {

for (sigma2.b in c(1,100)) {

for (sigma2.w in c(1,100)) {

for (sigma2.e in c(1,100)) {

table.mean=c();

j=j+1;

roh=sigma2.w/(sigma2.w+sigma2.e)

lambda=sigma2.b/(sigma2.b+(sigma2.w+sigma2.e)/i)

for (rep_time in 1:1000)   {

#Create true score theta

set.seed(rep_time*10+1+j*100000+400000000)

nu=rnorm(k, mean = 0, sd = sqrt(sigma2.b))

nu.ik=matrix(rep(nu,i),byrow=TRUE,nrow=i,ncol=k)

set.seed(rep_time*10+2+j*100000+400000000)

r2=matrix(rnorm(i*k, mean=0,sd=sqrt(sigma2.w)),nrow=i,ncol=k)

theta=nu.ik+r2

theta.mean=mean(theta)

theta.cluster.mean=colMeans(theta)
theta.Svar=var(theta.cluster.mean)/k
theta.cluster.var=sapply(1:k, function(y) var(theta[,y]))
theta.cluster.var.mean=mean(theta.cluster.var)
theta.cluster.mean.var=var(theta.cluster.mean)
theta.var=theta.cluster.mean.var+theta.cluster.var.mean
#Create observed score X
set.seed(rep_time*10+3+j*100000+400000000)
e=matrix(rnorm(i*k,mean=0,sd=sqrt(sigma2.e)),nrow=i,ncol=k)
e.cluster.mean=colMeans(e)
x=theta+e
e.mean=mean(e)
e.mean.var=var(e.cluster.mean)/k
e.var=var(as.numeric(e))
x.mean=mean(x)
x.cluster.mean=colMeans(x)
x.se=sqrt(var(x.cluster.mean)/k)
x.Svar=var(x.cluster.mean)/k
x.cluster.var=sapply(1:k, function(y) var(x[,y]))
x.cluster.var.mean=mean(x.cluster.var)
x.cluster.mean.var=var(x.cluster.mean)
x.var=x.cluster.mean.var+x.cluster.var.mean
set.seed(rep_time*10+0+j*100000+400000000)
h=rnorm(10,mean=0,sd=sqrt(x.Svar))
h.mean=mean(h)
h.var=var(h)
cor=cor(as.numeric(theta.cluster.mean), as.numeric(e.cluster.mean))
#Create imputation data sets
```

```
#Create random pick from empirial distribution of mu tilda
mu.tilda=x.mean+h
mu.tilda.mean=mean(mu.tilda)
set.seed(rep_time*10+4+j*100000+400000000)
g=matrix(rnorm(k*10,mean=0,sd=sqrt((1-lambda)*sigma2.b)),nrow=10,ncol=k)
g.mean=mean(g)
x.ikm=rep(x,10)
dim(x.ikm)=c(i,k,10)
nu.tilda.ikm=rep(0,i*k*10)
dim(nu.tilda.ikm)=c(i,k,10)

for (y in 1:10){
a=matrix(rep(g[y,],i),byrow=TRUE,nrow=i,ncol=k);
b=matrix(rep(lambda*x.cluster.mean,i),byrow=TRUE,nrow=i,ncol=k);
c=matrix(rep((1-lambda)*mu.tilda[y],i*k),nrow=i,ncol=k);
nu.tilda.ikm[,,y]=a+b+c;
}
nu.tilda.ikm.mean=mean(nu.tilda.ikm)
set.seed(rep_time*10+5+j*100000+400000000)
f.ikm=rnorm(i*k*10,mean=0,sd=sqrt((1-roh)*sigma2.w))
f.ikm.mean=mean(f.ikm)
dim(f.ikm)=c(i,k,10)
theta.tilda=roh*x.ikm+(1-roh)*nu.tilda.ikm+f.ikm

theta.tilda.mean=rep(0,10)
theta.tilda.cluster.mean=matrix(rep(0,k*10),nrow=10,ncol=k)
theta.tilda.Svar=rep(0,10)
theta.tilda.cluster.mean.var=rep(0,10)
theta.tilda.cluster.var=matrix(rep(0,k*10),nrow=10,ncol=k)
theta.tilda.cluster.var.mean=rep(0,10)
theta.tilda.cluster.mean.var=rep(0,10)
theta.tilda.var=rep(0,10)
for (y in 1:10) {
theta.tilda.mean[y]=mean(theta.tilda[,,y])
theta.tilda.cluster.mean[y,]=colMeans(theta.tilda[,,y])
theta.tilda.Svar[y]=var(theta.tilda.cluster.mean[y,])/k
theta.tilda.cluster.var[y,]=sapply(1:k,function(q) var(theta.tilda[,q,y]))
theta.tilda.cluster.var.mean[y]=mean(theta.tilda.cluster.var[y,])
theta.tilda.cluster.mean.var[y]=var(theta.tilda.cluster.mean[y,])
theta.tilda.var[y]=theta.tilda.cluster.var.mean[y]+theta.tilda.cluster.mean.var[y]
}
theta.tilda.SM=mean(theta.tilda.mean)
theta.tilda.UM=mean(theta.tilda.Svar)
theta.tilda.BM=var(theta.tilda.mean)
theta.tilda.VM=theta.tilda.UM+1.1*theta.tilda.BM
theta.tilda.var.mean=mean(theta.tilda.var)
table.mean=rbind(table.mean,t(c(
j,
sigma2.b,
sigma2.w,
sigma2.e,
k,
i,
theta.mean,
x.mean,
theta.tilda.SM,
g.mean,
nu.tilda.ikm.mean,
f.ikm.mean,
h.mean,
mu.tilda.mean,
h.var,
theta.var,
theta.cluster.mean.var,
theta.cluster.var.mean,
x.var,
x.cluster.mean.var,
x.cluster.var.mean,
theta.tilda.var.mean,
theta.Svar,
x.Svar,
theta.tilda.VM,
theta.tilda.UM,
theta.tilda.BM
)))
} # end of rep_time
colnames(table.mean)=c(
"j",
"sigma2.b",
"sigma2.w",
"sigma2.e",
"k",
"i",
"theta.mean",
"x.mean",
"theta.tilda.SM",
"g.mean",
"nu.tilda.ikm.mean",
"f.ikm.mean",
"h.mean",
"mu.tilda.mean",
"h.var",
"theta.var",
"theta.cluster.mean.var",
"theta.cluster.var.mean",
"x.var",
"x.cluster.mean.var",
"x.cluster.var.mean",
"theta.tilda.var.mean",
"theta.Svar",
"x.Svar",
"theta.tilda.VM",
"theta.tilda.UM",
"theta.tilda.BM"
)

write.csv(table.mean,file=
paste(paste("seed",j,"cluster",sigma2.b,sigma2.w,sigma2.e,"k",k,"i",i,sep="_"),"csv",sep="."),
row.names=FALSE)
theta.mth2.Svar=var(as.data.frame(table.mean)$theta.mean)
x.mth2.Svar=var(as.data.frame(table.mean)$x.mean)
SM.mth2.Svar=var(as.data.frame(table.mean)$theta.tilda.SM)
res=c(
colMeans(table.mean),
theta.mth2.Svar,
x.mth2.Svar,
SM.mth2.Svar)
table.final=rbind(table.final,res)
}#end of sigma2.e
}#end of sigma2.w
}#end of sigma2.b
}#end of i
}#end of k
colnames(table.final)=c(colnames(table.mean),"theta.mth2.Svar","x.mth2.Svar","SM.mth2.Svar")
write.csv(table.final,file="table_final_v100.csv",row.names=FALSE)
```

References

Bock, R. D., Mislevy, R., & Woodson, C. (1982). The next stage in educational assessment. *Educational Researcher, 11*(3), 4-11, 16.

Braun, H., Jenkins, F., & Grigg, W. (2006). *Comparing private schools and public schools using hierarchical linear modeling*. (NCES 2006-461). U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences. Washington, DC: U.S. Government Printing Office.

Cassel, C., Sa̎rndal, C.-E., & Wretman, J. H. K. (1977). *Foundations of inference in survey sampling*. New York: Wiley.

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Deming, W. E., & Stephan, F. F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association, 36*(213), 45-49.

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

U.S. Department of Education & National Center for Education Statistics. *The NAEP Guide*, (NCES 2000–456), by Horkay, N., editor. Washington, DC: U.S. Department of Education.

Elston, R. C., & Grizzle, J. E. (1962). Estimation of time-response curves and their confidence bands. *Biometrics, 18*(2), 148-159.

Frankel, M. R., & Frankel, L. R. (1987). Fifty years of survey sampling in the united states. *Public Opinion Quarterly – Supplement*: *50th Anniversary Issue*, 51(2), 127-138.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*, 2nd edn, London: Chapman & Hall/CRC.

Goldstein, H. (2010). *Multilevel statistical models*. New York: Wiley.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1993). *Sample survey methods and theory: Methods and applications*: New York: Wiley.

Kalton, G. (1983). *Introduction to survey sampling*. Beverly Hills: Sage Publications.

Kelley, T. L. (1923). *Statistical method.* New York: The Macmillan Company.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Lindman, H. R. (1974). *Analysis of variance in complex experimental designs*. San Francisco: Freeman.

Lord, F. M. (1959). Statistical inferences about true scores. *Psychometrika*, *24*(1), 1-17.

Lord, F. M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement*, *22*(2), 259-267.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mahalanobis, P. C. (1946). Sample surveys of crop yields in india. *Sankhyā: The Indian Journal of Statistics, 7*(3), 269-280.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*(3), 359-381.

Mislevy, R. J. (1989). *Foundations of a new test theory*. (ETS Research Report RR-89-52-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177-196.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*(4), 379-416.

Mislevy, R. J. (in press). On the proportion of "missing data" in classical test theory. *ETS Research Memorandum*. Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Chapter 3: Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics, 17*(2), 131-154.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, 97*(4), 558.

Rao, J. N. K. (2003). *Small area estimation*. New York: Wiley.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Beverly Hills: Sage Publications.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association, 72*(359), 538-543.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rust, K.F., Krenzke, T., Qian, J., and Johnson, E.G. (2001). Sample design for the national assessment. In J.E. Carlson and N.L. Allen (Eds.), *The NAEP 1998 technical report*. Washington, DC: National Center for Education Statistics.

Sirotnik, K., & Wellington, R. (1977). Incidence sampling: An integrated theory for “matrix sampling”. *Journal of Educational Measurement, 14*(4), 343-399.

Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of complex surveys*. New York: Wiley.

Spearman, C. (1904a). " General intelligence," objectively determined and measured. *The American journal of psychology*, *15*(2), 201-293.

Spearman, C. (1904b). The proof and measurement of association between two things. *The American journal of psychology, 15*(1), 72-101.

Spearman, C. (1907). Demonstration of formulæ for true measurement of correlation. *The American journal of psychology*, *18*(2), 161-169.

Spearman, C. (1913). Correlations of sums or differences. *British Journal of Psychology*, *5*(4), 417-426.