

Evaluating and Improving Missing-Data Handling in ERA-GEE: A Simulation and Empirical
Study with High-Dimensional Longitudinal Data

by

Bryce Phillips

A Thesis submitted to the Faculty of Graduate and Postdoctoral Studies of the University of
Manitoba in partial fulfilment of the requirements of the degree of

MASTER OF ARTS

Department of Psychology

University of Manitoba

Winnipeg

Copyright © 2026 by Bryce Phillips

Abstract

Extended Redundancy Analysis combined with Generalized Estimating Equations (ERA-GEE) provides a flexible framework for modeling high-dimensional longitudinal data by jointly summarizing complex predictor domains and accounting for correlated outcomes. A key limitation of ERA-GEE, however, is its reliance on a complete dataset for estimation. In practice, missing data are typically handled through preprocessing approaches such as listwise deletion (LWD) and multiple imputation (MI) prior to model fitting. Because these procedures are detached from the component construction, estimating-equation structure, and correlation assumptions that govern ERA-GEE, they can induce biased estimation and inefficient inference. This thesis proposes a least-squares imputation (LSI) strategy that is embedded directly within the ERA-GEE estimation procedure, yielding a unified framework for simultaneous parameter estimation and missing-data management. Rather than deleting or imputing observations prior to analysis, the proposed LSI-ERA-GEE approach iteratively updates model parameters and missing values within a single estimation criterion. A Monte Carlo simulation study evaluates the performance of LSI-ERA-GEE relative to LWD and MI across a range of missing-data conditions, including varying proportions of missingness and multiple missing mechanisms. Results show that across missing-data conditions, LSI-ERA-GEE reduces bias and improves estimation accuracy relative to competing approaches, particularly under low-to-moderate proportions of missingness. An empirical application using cognitive outcomes measured at three timepoints from the Canadian Longitudinal Study on Aging further demonstrates that parameter estimates, statistical significance, and predictive performance are highly sensitive to the choice of missing-management method. Overall, LSI-ERA-GEE produces the most precise estimates and lowest prediction error.

Keywords: missing-data management, Extended Redundancy Analysis-Generalized Estimating Equations, high-dimensional longitudinal data, Monte Carlo simulation, Canadian Longitudinal Study on Aging

Acknowledgements

I would like to acknowledge and sincerely thank my supervisor, Dr. Sunmee Kim, for her unwavering support, dedication, and mentorship throughout the completion of my thesis.

Whether it was providing written feedback on my thesis, organizing additional opportunities for my professional development, or sharing advice from her experiences in the academic world, Dr. Kim always went above and beyond as a supervisor. The guidance and leadership she provided was invaluable to me, and for that I thank her greatly.

I would also like to acknowledge and thank my committee members, Dr. Depeng Jiang and Dr. Richard Kruk. During the completion of my thesis, Dr. Jiang and Dr. Kruk were always willing to spare time from their busy schedules to share their knowledge and feedback, which allowed me to elevate the quality of my work. I am incredibly grateful for their time and the support they provided.

Lastly, I would like to acknowledge the funding sources that provided financial support for this thesis. Thank you to the Natural Sciences and Engineering Research Council of Canada (NSERC) for awarding me the Canadian Graduate Scholarship – Masters, the University of Manitoba for awarding me the University of Manitoba Graduate Fellowship, and the Department of Psychology for awarding me the Psychology Graduate Fellowship. I would also like to acknowledge additional support that was provided from Dr. Kim's research funding. I am extremely grateful for all the support provided, as the completion of this thesis would not have been possible without the financial support from these sources.

Table of Contents

Abstract.....	ii
Acknowledgements	iv
List of Tables.....	vi
List of Figures.....	vii
Evaluating and Improving Missing-Data Handling in ERA-GEE: A Simulation and Empirical Study with High-Dimensional Longitudinal Data	1
Analytic Context and Limitations of Standard Longitudinal Models	1
ERA-GEE for High-Dimensional Longitudinal Data	4
The Challenge of Missing Data in ERA-GEE	6
Existing Approaches to Missing Data	7
Listwise Deletion	7
Multiple Imputation	8
Toward an Integrated Missing-Management Method for ERA-GEE	9
Evaluating Performance Under Different Missing Mechanisms	12
Objective	13
Methods.....	15
ERA-GEE.....	15
Model Specification.....	15
Parameter Estimation and Significance Testing	17
Conventional Missing-Management Methods	18
LWD	18
MICE	19
LSI Embedded in ERA-GEE.....	20
Simulation Study Design.....	22
Model Specification and Data Generation.....	22
Missing-Data Design and Implementation	23
Experimental Conditions	25
Performance Metrics	26
Empirical Application	29
Data.....	29
Variables	29

Model Specification.....	29
Missing-Data Exploration.....	30
Evaluation of Missing-Management Methods	32
Results	33
Simulation Results.....	33
Method Feasibility	33
Estimator Behaviour	35
Empirical Application Results.....	38
Missing-Data Analysis.....	38
Bootstrap Convergence Behaviour	39
Estimator Comparison	40
Prediction Accuracy	42
Substantive Results from the LSI-ERA-GEE Model	42
Discussion	43
LWD	43
MICE	45
LSI-ERA-GEE	46
Limitations	48
Future Directions	51
References	53
Appendix A	70
Appendix B	74
Appendix C	79

List of Tables

Table 1: CLSA Empirical Application Variable Measures and Descriptive Statistics	61
Table 2: Empirical Application Within-Predictor-Block Correlations.....	62
Table 3: Empirical Application ERA-GEE Analysis Results	63
Table B1: Convergence Probabilities Across Missing-Data Conditions Under LWD.....	74
Table B2: Proportion of Retained Sample Across Missing-Data Conditions Following LWD	

.....	75
Table B3: Mean Convergence Time in All Simulation Conditions	76
Table B4: Mean Number of Iterations to Convergence in All Simulation Conditions	77
Table B5: Mean Total-Loss Value at Convergence in All Simulation Conditions.....	78
Table C1: Variance Results: MCAR Simulation Conditions	79
Table C2: Variance Results: MAR Simulation Conditions	80
Table C3: Variance Results: MNAR Simulation Conditions	81
Table C4: Bias Results: MCAR Simulation Conditions	82
Table C5: Bias Results: MAR Simulation Conditions.....	83
Table C6: Bias Results: MNAR Simulation Conditions.....	84
Table C7: RMSE Results: MCAR Simulation Conditions	85
Table C8: RMSE Results: MAR Simulation Conditions	86
Table C9: RMSE Results: MNAR Simulation Conditions	87

List of Figures

Figure 1: Schematic Diagram of ERA-GEE Model Used for Simulated Data Generation	
.....	65
Figure 2: Regression Coefficient Bias Across Missing-Management Methods under MAR and MNAR.....	66
Figure 3: Bootstrapped Component Weight Distributions Across Missing-Management Methods	
.....	67
Figure 4: Bootstrapped Regression Coefficient Distributions Across Missing-Management Methods.....	68

Figure 5: Schematic of ERA-GEE Model from the Empirical Application with LSI-ERA-GEE

Results 69

Evaluating and Improving Missing-Data Handling in ERA-GEE: A Simulation and Empirical Study with High-Dimensional Longitudinal Data

Analytic Context and Limitations of Standard Longitudinal Models

High-dimensional longitudinal panel studies collect information on many variables over a select number of assessments that are usually spaced out by a few years. The aim of these studies is often to investigate phenomena with long-term trajectories while examining the impact of high-dimensional sets of variables on the change (e.g., Ekström et al., 2024; Meister & Zahodne, 2022; Park et al., 2003). For example, every three years, the Canadian Longitudinal Study on Aging (CLSA) collects data on thousands of variables from a sample of approximately 50,000 aging Canadians. With an increase in the aging population (Statistics Canada, 2022), studies such as the CLSA are important as they allow researchers to investigate the long-term process of aging and what factors impact this process. By incorporating many variables measured repeatedly over time, high-dimensional longitudinal studies enable more comprehensive modeling of complex, time-varying processes.

From an analytic standpoint, two fundamental features of high-dimensional longitudinal data must be addressed when selecting an appropriate statistical modeling framework. First, the inclusion of high-dimensional variables as potential predictors or covariates—often encompassing demographic, behavioural, physiological, and psychological measures—poses challenges such as overfitting, multicollinearity, and interpretability issues (Otse et al., 2025; Rahnenführer et al., 2023). In a statistical sense, the term high dimensionality refers to settings where the number of variables p is large relative to the number of observations n , such that the effective rank of the design matrix is low and classical estimation becomes unstable. This condition arises across a wide range of empirical domains, from genomics and neuroimaging to

the behavioural and social sciences. In social science research, high-dimensional challenges can emerge even when p is not formally larger than n , because variables measuring psychological, health, and social constructs are often substantially intercorrelated and conceptually overlapping. Such intercorrelations can reduce the effective information contained in the design matrix, producing estimation instability and interpretability challenges similar to those encountered in conventionally high-dimensional settings.

Second, repeated measurements within individuals from high-dimensional longitudinal data introduce within-subject correlations, which must be properly modeled to ensure valid statistical inferences (Kenny & Judd, 1986). For example, if an individual's cognitive functioning is assessed at two time points spaced 2 months apart, the responses are likely to be correlated due to underlying subject-specific factors. This intra-individual dependence violates the assumption of independent observations and requires models that explicitly incorporate correlation structures. Therefore, effective modeling approaches for high-dimensional longitudinal data must address both the complexity introduced by the large number of predictors and the correlation structure.

While numerous methods have been developed to model correlated data, Generalized Estimating Equations (GEE) and linear multilevel models remain among the most widely used approaches (e.g., Guo et al., 2024; Kirkland et al., 2023; Quick et al., 2022; Raina et al., 2021). These methods are particularly effective for addressing within-subject correlations that arise from clustered or repeated observations, such as those seen in longitudinal, nested, or panel data. GEE provides robust estimation of population-averaged effects and is relatively insensitive to misspecification of the correlation structure (Kim et al., 2020; Lee et al., 2019; L. Wang et al., 2012), making it well-suited for marginal inference—that is, drawing conclusions about averaged

effects across the population rather than individual-specific trends. In contrast, linear multilevel models (also known as mixed-effects models or hierarchical linear models) estimate subject-specific effects by modeling variation across units through random effects. By including random intercepts and, when appropriate, random slopes, these models capture unit-specific deviations allowing for individual differences in baseline levels and change over time or across conditions.

However, standard GEE and linear multilevel models are not well-equipped for high-dimensional settings where variable selection and dimensionality reduction become essential for both model interpretability and computational feasibility (Holmes Finch, 2018; L. Wang et al., 2012). First, both approaches rely on maximum likelihood or quasi-likelihood estimation methods that become increasingly unstable and computationally intensive as dimensionality grows. Estimation may suffer from convergence issues, inflated standard errors, or numerical instability (Yoo et al., 2014)—especially when multicollinearity among predictors renders the design matrix nearly singular. These challenges are further amplified in multilevel models due to the expansion of the random effects structure. As the number of predictors grows, models may require multiple random slopes in addition to random intercepts, substantially increasing the number of parameters and the complexity of the covariance structure. Parameter estimation in this setting involves intensive matrix operations, including repeated inversions of large covariance matrices. Iterative methods such as restricted maximum likelihood, Expectation-Maximization, and Bayesian approaches using Markov Chain Monte Carlo can become prohibitively slow or unstable in this context, limiting their practicality for large-scale studies with many variables and participants.

Furthermore, neither GEE nor standard multilevel models incorporate built-in mechanisms for regularization or variable selection, leaving them vulnerable to two key

challenges of high-dimensional data. First, in high-dimensional contexts many predictors can be weakly associated with the outcome which can lead to overfitting. Second, multicollinearity from correlations between predictors can inflate standard errors (Vatcheva et al., 2016). When unaccounted for, these challenges not only compromise the generalizability and stability of the model, but also obscure meaningful patterns and impede interpretability. While regularized extensions such as penalized GEE and shrinkage-based multilevel models have been proposed to mitigate these issues (e.g., Holmes Finch, 2018; L. Wang et al., 2012), their application in psychological and health research remains limited. Contributing factors include computational burden, sparse documentation, and limited accessibility in mainstream software. These limitations underscore the need to explore alternative frameworks that are purposefully designed for the structure and demands of high-dimensional longitudinal data.

ERA-GEE for High-Dimensional Longitudinal Data

To address the challenges posed by high-dimensional and correlated data structures, recent methodological advancements have combined dimension reduction techniques with longitudinal modeling frameworks capable of handling within-unit correlations. One such development is the novel integration of Extended Redundancy Analysis (ERA; Takane & Hwang, 2005) with GEE (Liang & Zeger, 1986)—referred to as ERA-GEE (Kim et al., 2020). This approach combines the interpretability and dimension reduction capabilities of ERA—particularly well-suited for scenarios where a large set of potentially correlated predictors must be considered—with GEE’s capacity to account for within-subject or within-cluster correlation in repeated measures or nested data. Importantly, ERA-GEE incorporates ridge-type regularization directly into its parameter estimation procedure, mitigating the risk of overfitting, improving estimation reliability, and reducing sensitivity to multicollinearity. By uniting these

complementary strengths, ERA-GEE offers a flexible and interpretable solution for modeling complex longitudinal or clustered data with many predictors.

In practice, researchers often try to address the previously mentioned challenges of high-dimensional settings by excluding variables or aggregating them into summary scores (e.g., through averaging or ad hoc weighting). While such strategies may simplify analysis, they can also obscure the unique contributions of individual predictors and introduce new interpretability concerns. Rather than excluding predictors or treating all variables equally, ERA identifies linear combinations of predictors—components—that are optimized to explain variance in the outcome variable(s), thereby reducing the number of parameters to interpret (Takane & Hwang, 2005). Moreover, ERA allows researchers to predefine conceptually meaningful groups of predictors—for example, by combining education, income, savings, and related variables into a socioeconomic status (SES) component—offering a principled way to retain theoretical structure while reducing model complexity. This flexibility makes ERA especially valuable in applied psychological and health research, where predictors frequently reflect overlapping constructs grounded in theory. By capturing shared variance within these clusters and preserving relevance to the outcome, ERA facilitates more interpretable and efficient modeling in high-dimensional contexts.

While ERA offers a powerful framework for interpretable dimension reduction, its conventional form is limited in handling multiple outcomes with non-normal and correlated structures—common features in longitudinal and clustered data. Its original formulation accommodates multiple outcomes only under the assumption of multivariate normality, which is rarely appropriate in applied contexts. To address this limitation, ERA-GEE integrates GEE into the ERA framework (Kim et al., 2020). By modeling intra-individual correlation through a

working correlation matrix, GEE avoids the need for full specification of subject-specific trajectories, focusing instead on marginal population-level trends. Moreover, it supports the analysis of multiple outcome variables with differing distributions (e.g., binary disease indicators, count-based behaviours, and continuous cognitive scores), which cannot be jointly modeled under standard multivariate normality. In this thesis, the term multivariate is used to refer broadly to two common scenarios: (a) multiple distinct outcomes, which may share the same or have different distributions, and (b) repeated measures of a single outcome, which yield a series of correlated responses of the same type over time. As such, ERA-GEE provides a flexible and interpretable modeling solution for high-dimensional, multivariate longitudinal data, including large-scale studies like the CLSA, frequently used in psychological and health research.

The Challenge of Missing Data in ERA-GEE

Despite its methodological strengths, the practical application of ERA-GEE is currently constrained by a major limitation: it requires a complete dataset—that is, no missing observations across any of the variables involved—for estimation and statistical inference. In practice, this is rarely achievable, especially in large-scale aging studies like the CLSA, where missing data are prevalent and often non-random. As individuals age, follow-up assessments may be missed due to deteriorating health, mobility challenges, cognitive decline, or attrition from mortality—leading to complex patterns of dropout that vary across time points and demographic subgroups (e.g., Brilleman et al., 2010; Chatfield et al., 2005; Feng et al., 2012). These mechanisms can violate the assumptions underlying many conventional approaches to handling missing data, as will be discussed in the following sections.

Existing Approaches to Missing Data

Listwise Deletion

Currently, the most commonly used approach for managing missing data within ERA-GEE is listwise deletion (LWD), also known as complete-case analysis. This method excludes from the analysis any individual with a missing value on any of the variables involved. While straight forward to implement, LWD carries serious consequences in high-dimensional contexts—such as those encountered in ERA-GEE. In large-scale studies like the CLSA, LWD can result in a loss of over 50% of the original sample as missingness often occurs across many variables (Feng et al., 2012). Such substantial reductions in sample size severely diminish statistical power, making it more difficult to detect true effects when they exist (Feng et al., 2012; Graham, 2009). Moreover, the increased sampling error associated with smaller analytic samples leads to less precise estimates, ultimately weakening the reliability and validity of the model's conclusions.

More fundamentally, the validity of LWD depends on the strong assumption that the data are missing completely at random (MCAR), meaning the probability of a value being missing is unrelated to both observed and unobserved values (Rubin, 1987). As illustrated by the dropout mechanisms described above, this assumption is rarely satisfied (Graham, 2009). When the MCAR assumption is violated, LWD can lead to biased parameter estimates and limit the generalizability of findings, as the retained sample no longer reflects the characteristics of the full population (Feng et al., 2012; Graham, 2009). Consequently, relying on LWD for missing-data management in ERA-GEE analyses risks producing biased estimates and unreliable inference, thereby limiting its utility in real-world longitudinal research.

Multiple Imputation

Multiple imputation (MI) is one of the most widely used approaches for handling missing data across various statistical frameworks. In longitudinal research, the most common implementation is MI with fully conditional specification, also known as multiple imputation by chained equations (MICE). In general, MI involves replacing each missing value with multiple plausible estimates derived from the observed data, resulting in several complete datasets—typically 30 or more (Azur et al., 2011; Woods et al., 2024).

Under MICE, missing values are imputed sequentially on a variable-by-variable basis using iterative univariate models tailored to each variable's distribution (e.g., linear regression for continuous variables and logistic regression for binary variables; Azur et al., 2011; van Buuren & Groothuis-Oudshoorn, 2011). This framework accommodates datasets with mixed variable types by specifying a separate imputation model for each incomplete variable, treating it as the outcome, and using all other variables as predictors. Once imputed datasets are generated, each is analyzed using the intended statistical model, and the results from each dataset are combined using Rubin's rules. Specifically, the pooled estimate is the average of each parameter estimate calculated across all imputed datasets, and the total variance accounts for both within- and between-imputation variability (Rubin, 1987). By generating many possible parameter estimates and taking the average of them, this pooling procedure yields final point estimates and standard errors that reflect uncertainty due to missingness.

MI is highly versatile and has been successfully applied in a range of frameworks, including GEE, linear multilevel models, and generalized linear models (e.g., Ditlhong et al., 2018; Gupta & Grover, 2017; Huque et al., 2018). Furthermore, MI has also been shown to produce less biased estimates and more accurate standard errors compared to LWD (e.g., Ji et al.,

2018; Steif et al., 2022). Despite its versatility and advantages over LWD, to date, no prior work has systematically evaluated the performance of MI within the ERA-GEE framework. Given the growing use of ERA-GEE for modeling complex, correlated outcomes and the pervasive issue of missing data in longitudinal research, it is necessary to examine whether MI provides a feasible and effective approach for managing missing data in this context.

Although MI is widely regarded as a flexible, general-purpose approach, two limitations are particularly relevant to ERA-GEE. The first concerns the lack of alignment between the imputation model and the analysis model. MI is implemented as a preprocessing step and imputes each variable independently, without incorporating the structural relationships imposed by the analysis model. In contrast, ERA-GEE performs supervised dimension reduction by constructing components from predefined blocks of predictors, aiming to maximize the explained variance in the outcome variables (Kim et al., 2020). Because MICE does not account for this block structure during imputation, it may distort the underlying relationships among predictors that are critical for component construction. This misalignment undermines the assumption of congeniality—namely, that the imputation and analysis models are conceptually compatible—which is essential for valid inference under the MI framework (Takahashi, 2017).

The second limitation arises from the behaviour of MI in high-dimensional settings, which are typical in ERA-GEE applications. In the MICE implementation of MI, as the number of variables in an analysis grows, the conditional models used for imputation become difficult to estimate reliably, particularly in the presence of multicollinearity (Collins et al., 2001; Vatcheva et al., 2016), leading to instability and potential overfitting (van Buuren & Groothuis-Oudshoorn, 2011). These challenges are amplified by the structural characteristics of ERA-GEE, where predictors within each block are intentionally selected to be conceptually related and often highly

correlated to facilitate interpretable component construction. As a result, the imputation step may introduce distortions into the joint distribution of predictors, rather than simply recovering missing information. Consequently, the reliability of model estimates and the interpretability of component structures can become uncertain.

Toward an Integrated Missing-Management Method for ERA-GEE

To address the aforementioned limitations, this thesis proposes a novel extension of the ERA-GEE framework that incorporates a built-in least-squares imputation (LSI) procedure into the estimation algorithm, enabling joint estimation of model parameters and missing values. LSI is an integration-based missing-management approach in which missing values are treated as additional unknown quantities and estimated jointly with model parameters in a single iterative process (Gabriel & Zamir, 1979; Gifi, 1990; Mair & de Leeuw, 2009; Wasito & Mirkin, 2005). In least-squares-based models, this can be achieved by embedding missing values directly into the model's objective function. The estimation procedure begins with initial placeholder values for the missing entries and then alternates between updating model parameters and imputing missing values until convergence is achieved. Therefore, in LSI, both missing values and model parameters are estimated by minimizing a common least-squares objective function.

This proposed integrated approach to missing-data management within ERA-GEE is motivated by LSI's successful application to Generalized Structured Component Analysis (GSCA; Hwang & Takane, 2015), which is a component-based modeling framework that shares key algorithmic features with ERA and ERA-GEE. GSCA is a more general model commonly used to examine relationships among latent constructs, and under certain constraints, a simplified GSCA model becomes equivalent to ERA—making ERA a special case of GSCA (Hwang & Takane, 2015). GSCA and ERA share essential algorithmic characteristics that render the

integration of LSI into ERA-GEE plausible. From an algorithmic standpoint, both methods employ alternating least-squares (ALS) procedures that iteratively minimize a global loss function. Specifically, a least-squares objective function is repeatedly minimized across iterative steps, where parameter estimates are updated sequentially while holding others fixed (Hwang & Takane, 2015; Takane & Hwang, 2005). This iterative optimization continues until convergence is achieved, enabling simultaneous estimation of components and model parameters. In the proposed ERA-GEE extension, an additional step is added to this iterative procedure where missing values are also estimated within the same least-squares procedure used for estimating model parameters.

By embedding this LSI procedure directly into ERA-GEE's iterative estimation routine, the proposed extension addresses the previously outlined limitations of MICE. First, because missing values are imputed as part of the same optimization process used to estimate components, the imputation inherently respects the grouped structure of predictors and the relationships required for component formation—resolving the issue of model incongruence. Second, since LSI does not rely on separate univariate regression models for each variable, it avoids the high-dimensional instability, multicollinearity, and overfitting concerns that often affect MICE. Furthermore, the joint estimation of components, model parameters, and missing values eliminates the need to pool results across multiple imputed datasets, thereby simplifying inferences and improving computational efficiency.

To my knowledge, this research is the first to extend ERA-GEE with a built-in missing-data estimation step, resulting in a fully integrated model that performs simultaneous component extraction, regression modeling of correlated outcomes, and imputation of missing values. The proposed method—titled LSI-ERA-GEE—is expected to not only improve the applicability of

ERA-GEE for real-world longitudinal datasets (such as the CLSA) but will also offer a computationally efficient and interpretable alternative to traditional imputation methods.

Therefore, the contributions of this thesis are expected to be both methodological and practical, with implications for a broad range of fields involving high-dimensional predictors and longitudinal outcome structures.

Evaluating Performance Under Different Missing Mechanisms

As the proposed method introduces an integrated imputation strategy within ERA-GEE, understanding the theoretical underpinnings of missing-data mechanisms, which refer to the causes of missing data, is critical for its evaluation and justification. As with other missing management approaches, the performance of LSI-ERA-GEE is expected to vary depending on the underlying missing mechanism (e.g., Chen et al., 2023; Dittlhong et al., 2018; Ji et al., 2018). When evaluating missing-management methods, the goal is to understand how a method will perform under various missing conditions that occur in real-world analyses. This is an important consideration, as different missing mechanisms have distinct impacts on model performance, even with a constant proportion of missing data (e.g., Chen et al., 2023; Ji et al., 2018). Moreover, MI and LSI both assume that the underlying missing mechanism is missing at random (MAR) or MCAR, both of which are described in the following paragraph. It is therefore important to validate that the methods are effective when this assumption holds and examine if they are robust to violations of this assumption.

Rubin (1976) categorized missing mechanisms into three distinct types: MCAR, MAR, and missing not at random (MNAR), each of which can occur in longitudinal data. Data is said to be MCAR when the probability of a value being missing is completely random. For example, a value is missing in an outcome variable, and the probability of missingness shares no correlation

with observed predictor values, nor the unobserved value itself. Data is said to be MAR when the probability of a value being missing is related to other observed variables, but not the unobserved value itself. For example, a case has a missing value in an outcome variable, and the probability of missingness is correlated with an observed predictor variable, but not the unobserved value itself. Data is said to be MNAR when the probability of a value being missing is exclusively related to the unobserved value itself. For example, a case has a missing value in an outcome variable, and the probability of missingness is correlated with the unobserved value itself.

Real-world analyses cannot be used to evaluate these issues as it is not possible to verify the underlying missing mechanism in empirical datasets due to the missing values' unobserved nature (Graham, 2009). Therefore, simulation studies are essential as they allow for the direct manipulation of missing mechanisms, as well as other missing-data factors, supporting the recreation of missing-data conditions that can be anticipated in real-world analyses. To investigate the impact of relevant missing-data factors, simulation-based evaluations were utilized in the present study. This allowed for assessment of the bias and efficiency of parameter estimates under different missing conditions, and the comparison of the performance of various missing management strategies—such as LWD, MICE, and LSI-ERA-GEE.

Objective

The overarching goal of this thesis was to examine the impact of missing data on parameter estimation in the ERA-GEE framework and to develop a methodologically sound and computationally efficient solution for addressing this challenge—an area that remains underexplored in the existing literature. Rather than simply documenting the consequences of missingness, this thesis introduces and evaluates a novel extension of ERA-GEE that integrates an LSI procedure directly into its alternating estimation algorithm. This built-in approach was

designed to jointly estimate model parameters and missing values in a unified process, thereby preserving the predictor-component structure central to ERA, and addressing limitations inherent in conventional preprocessing-based imputation methods.

To assess the relative strengths and weaknesses of existing and proposed methods, this study compared three approaches to handling missing data: (1) LWD, a widely used but potentially inefficient deletion-based method, (2) MICE, a flexible and general-purpose external strategy, and (3) the newly proposed LSI-ERA-GEE model. These approaches were evaluated through an extensive Monte Carlo simulation study that systematically varied missing data mechanisms, proportions of missing data, and sample sizes. In each missing-data scenario, all three missing-management methods were applied to evaluate and compare each method's feasibility and estimation performance.

To complement the simulation findings and demonstrate the practical relevance of the proposed approach, all three methods were also applied in an empirical analysis using real-world data from the CLSA. Despite its complexity and scale, most existing studies using CLSA data have relied on ad hoc or simplistic missing-data strategies. This thesis provides the first systematic comparison of missing-data-management techniques within the ERA-GEE framework using CLSA data, offering insights into how these methods perform under realistic high-dimensional longitudinal conditions. Furthermore, directly comparing the results following the application of each missing-management method clearly demonstrates how a researcher's choice of missing-management method can impact results. Together, the methodological development, simulation-based validation, and empirical application provide a strong foundation for the theoretical and practical contributions of the proposed method.

Methods

ERA-GEE

Model Specification

Within an ERA-GEE model we have Q outcome variables and K unique sets of predictors, where each set consists of P_k predictors ($k = 1, \dots, K$) and $P = \sum_{k=1}^K P_k$. Let y_{iq} denote the value of the q th variable for the i th respondent ($i = 1, \dots, N$; $q = 1, \dots, Q$). We assume that y_{iq} follows an exponential family distribution with a mean μ_{iq} and variance $\phi\sigma_{iq}^2$, where ϕ is a dispersion parameter that may not be of primary substantive interest. Let x_{ikp} denote the p th predictor within the k th predictor block for individual i , and let w_{kp} denote the corresponding component weight. Let the component score, which is the sum of weighted predictors for the i th individual on the k th component be denoted by $f_{ik} = \sum_{p=1}^{P_k} x_{ikp}w_{kp}$. Let β_{kq} denote the regression coefficient relating the k th component to the q th outcome variable. Let η_{iq} denote the linear predictors for the q th outcome for individual i , and $g(\cdot)$ denote a known link function. Within an ERA-GEE model, we assume all variables are standardized with zero means and unit variances (Takane & Hwang, 2005). The formula for the ERA-GEE model is then expressed as

$$g(\mu_{iq}) = \eta_{iq} = \sum_{k=1}^K \left[\sum_{p=1}^{P_k} x_{ikp}w_{kp} \right] \beta_{kq} = \sum_{k=1}^K f_{ik}\beta_{kq},$$

such that the marginal expectation of the outcome μ_{iq} is related to a linear predictor through the specified link function.

Let $\tilde{y}_i = [y_{i1}, \dots, y_{iQ}]'$ be a Q by 1 vector of outcomes for individual i . Let Σ_i be the Q by Q within-individual covariance matrix of $\tilde{y}_i$. In standard ERA, it is assumed that outcome variables are independent, however, in ERA-GEE when multiple outcomes are measured

simultaneously for each individual, this assumption can be violated. Additionally, the true covariance structure is often unknown. To accommodate this dependence, ERA-GEE adopts the GEE framework to model the covariance structure using a working correlation matrix (Liang & Zeger, 1986). The covariance matrix of $\tilde{y}_i$ is specified as

$$\text{cov}(\tilde{y}_i) = \Sigma_i = \phi A_i^{\frac{1}{2}} R_i(\mathbf{a}) A_i^{1/2},$$

where $R_i(\mathbf{a})$ is a Q by Q working correlation matrix that is assumed to be fully specified by the vector of unknown nuisance parameters $\mathbf{a}$, and $A_i^{1/2}$ which is a Q by Q diagonal matrix of marginal variances with $\text{var}(\mu_{iq})$ as the q th diagonal element (Liang & Zeger, 1986).

The interpretation of the working correlation matrix depends on the nature of the correlated outcomes. In a multivariate outcome setting, the outcomes represent different constructs or response domains measured for the same individual, and the working correlation matrix captures their co-occurrence. In a longitudinal repeated-measures setting, the outcomes instead represent the same construct measured across time, and the working correlation matrix captures the temporal dependency among measurement occasions. Thus, although the same ERA-GEE formulation can accommodate both types of correlated outcomes, the specification of the working correlation structure should be guided by the substantive and measurement structure of the outcomes being analyzed. For example, several structures have been proposed for the working correlation matrix, including independence, exchangeable, and first-order autoregressive forms, and researchers should specify the structure they believe best represents the underlying relationship between all outcomes. However, exact specification of the working correlation structure is not required for consistent estimation of the mean-model parameters, provided that the mean model is correctly specified. The working correlation structure can therefore be treated as a nuisance component that primarily affects estimation efficiency, with

robust sandwich standard errors used to support valid inference under correlation misspecification.

Parameter Estimation and Significance Testing

While the general formulation of ERA-GEE allows for arbitrary link functions, the estimation procedure implemented here is based on a quadratic objective defined by the observed outcome scale. This corresponds to an identity-link specification and can be viewed as a Gaussian working approximation for the purposes of parameter estimation and missing-data updating.

An ERA-GEE model estimates three sets of unknown parameters—component weights, regression coefficients, and correlation parameters—using an iterative optimization procedure. Parameter estimation is carried out by minimizing the following penalized least-squares criterion:

$$\phi_{(a, \mathbf{W}, \mathbf{B})} = \sum_{i=1}^N [(\tilde{z}_i - \mathbf{B}'\mathbf{W}'\tilde{x}_i)' \Sigma_i^{-1} (\tilde{z}_i - \mathbf{B}'\mathbf{W}'\tilde{x}_i)] + \lambda_{\mathbf{W}} \text{trace}(\mathbf{W}'\mathbf{W}) + \lambda_{\mathbf{B}} \text{trace}(\mathbf{B}'\mathbf{B}),$$

where $\mathbf{B}$ is the K by Q matrix of regression coefficients, $\mathbf{W}$ is the P by K matrix of component weights, and $\lambda_{\mathbf{W}}$ and $\lambda_{\mathbf{B}}$ are the tuning parameters used to control the strength of the regularization applied to the component weights and regression coefficients, respectively. In general, the optimal values of these tuning parameters can be selected using G-fold cross-validation.

To minimize ERA-GEE's objective function, estimation proceeds using a regularized ALS algorithm. In this procedure, the overall objective function is minimized by updating each set of parameters in turn while holding the other parameters fixed. Each update step minimizes the same objective function with respect to one set of parameters at a time. A detailed description of the parameter estimation procedure used in ERA-GEE is described in Kim et al. (2020). The

procedure iterates through these updates until the convergence criterion is satisfied or until a predefined maximum number of iterations is reached. In practice, the algorithm typically converges reliably, although convergence within the maximum number of iterations is not guaranteed.

Statistical inference for the estimated parameters can be obtained using nonparametric resampling methods such as bootstrapping (Efron & Tibshirani, 1986). By repeatedly generating bootstrap samples and fitting the ERA-GEE model to each sample, an empirical distribution of parameter estimates can be obtained. These bootstrap distributions allow for the estimation of standard errors and the construction of confidence intervals.

Conventional Missing-Management Methods

LWD

LWD was implemented in both the simulation study and empirical application by incorporating an additional preprocessing step within the ERA-GEE procedure. In this step, all rows within the dataset containing missing values on any variables were removed prior to model estimation. After removing these rows, the remaining data were processed using the standard ERA-GEE estimation procedure, including variable scaling, parameter initialization, and parameter estimation. Under simulation conditions with high proportions of missing values and small sample sizes, the remaining sample after LWD occasionally became too small for model estimation. In such cases, the results were recorded as non-convergence events with parameter estimates stored as NA.

MICE

MICE was implemented using the *mice* package, version 3.19.0 (van Buuren & Groothuis-Oudshoorn, 2025). The imputation procedure was incorporated into the ERA-GEE function by performing imputation prior to variable scaling, and pooling parameter estimates after model estimation across imputed datasets.

The imputation procedure begins with specifying the imputation prediction matrix and the vector of imputation methods. The imputation predictor matrix is a $(P+Q)$ by $(P+Q)$ matrix of zeros and ones, indicating which variables are used as predictors in each imputation model. Rows correspond to variables being imputed, and columns correspond to potential predictors. In both the simulation and empirical analysis, missing values were present in all variables, and all variables were used in each imputation model. Consequently, the predictor matrix consisted of ones with zeros along the diagonal to exclude the variable currently being imputed from its own imputation model.

The imputation methods were specified using a vector of length $P+Q$, where each element indicates the imputation method used for the corresponding variable. As all variables in both analyses were normally distributed, the imputation method *norm.nob* was used for all variables. In the *mice* package, this corresponds to a conditional mean imputation model in which each variable with missing values is imputed using a linear regression model based on the predictors specified in the predictor matrix.

After specifying the predictor matrix and imputation methods, 30 imputed datasets were generated using the MICE algorithm. The procedure begins by initializing missing values using randomly sampled observed values from the same variable. Conditional imputation models are then sequentially fit for each variable, and missing values are updated accordingly. This iterative

process was repeated for five iterations, which is the default setting in the *mice* package and is generally sufficient for convergence of the imputed values (van Buuren & Groothuis-Oudshoorn, 2011).

Although the MICE procedure was integrated within the ERA-GEE function for computational convenience, the imputation itself was conducted independently of the ERA-GEE estimation algorithm. Once all 30 imputed datasets were generated, an ERA-GEE model was fit separately to each dataset. The resulting parameter estimates were then pooled across imputations using Rubin's rules (Rubin, 1987).

LSI Embedded in ERA-GEE

Conventional missing-management methods such as LWD or MICE treat missing values as a preprocessing problem that is addressed prior to model estimation. In contrast, the proposed LSI approach incorporates the imputation of missing values directly within the ERA-GEE estimation procedure. By embedding an imputation phase within the optimization algorithm, missing values are updated in a manner that is consistent with the ERA-GEE objective function. As a result, the imputed values and model parameters are estimated jointly within a unified estimation framework.

To achieve this integration, the standard ALS algorithm used in ERA-GEE is extended to include two sequential phases within each iteration. Let $\mathbf{X}$ denote the N by P predictor matrix and $\mathbf{Y}$ the N by Q outcome matrix. Let $\mathbf{X}_{ind}$ and $\mathbf{Y}_{ind}$ denote indicator matrices identifying missing entries in $\mathbf{X}$ and $\mathbf{Y}$, respectively. As an initialization step, missing values in both $\mathbf{X}$ and $\mathbf{Y}$ are replaced using column means, and all variables are standardized. Initial values for the component weights, regression coefficients, and correlation parameters are then specified.

At each iteration, the algorithm proceeds by alternating between parameter-updating and missing-value-updating steps. First, given the current (partially imputed) data matrices, the ERA-GEE objective function is minimized with respect to the model parameters. Specifically, the regression coefficients, component weights, and correlation parameters are updated sequentially, holding all other quantities fixed, following the standard ALS procedure.

Second, given the updated parameter estimates, missing values in $\mathbf{Y}$ and $\mathbf{X}$ are updated by minimizing the same least-squares objective function with respect to the missing elements. Because predictors and outcomes enter the model differently, separate subproblems are solved for outcomes and predictors. For each outcome variable j containing missing values, the least-squares objective is minimized with respect to the missing elements y_{ij} , yielding:

$$\hat{y}_{ij} = \mu_{ij} + \Sigma_{j-j} \Sigma_{-j-j}^{-1} (y_{i-j} - \mu_{i-j}),$$

where y_{i-j} denotes the remaining observed outcomes for individual i excluding variable j , and Σ is the Q by Q covariance matrix for outcomes which is assumed to be constant across all observations.

Similarly, for each predictor variable j containing missing values, the least-squares objective is minimized with respect to the missing predictor elements, yielding:

$$\hat{x}_{ij} = \frac{\mathbf{M}_j \Sigma^{-1} (y_i - x_{i-j} \mathbf{M}_{-j})'}{\mathbf{M}_j \Sigma^{-1} \mathbf{M}_j'},$$

where $\mathbf{M} = \mathbf{W}\mathbf{B}$ is the P by Q matrix linking predictors to outcomes, $\mathbf{M}_j$ denotes the row corresponding to predictor j , and x_{i-j} denotes the remaining predictors for individual i . Full derivations of these estimating equations are provided in Appendix A.

Following the update of model parameters and missing values, the data matrices are rescaled, component scores $\mathbf{F} = \mathbf{X}\mathbf{W}$ are recomputed, and the covariance matrix Σ is updated.

These steps define one full iteration of the extended ALS algorithm. The algorithm iterates through these updates until convergence.

Simulation Study Design

Model Specification and Data Generation

The simulation design was constructed to reflect the key structural features of the ERA-GEE framework, including predictor blocks defining components and correlated multivariate outcomes. Data were generated from an ERA-GEE model with two components, three predictors per component, and three outcomes. Figure 1 displays a schematic diagram of the ERA-GEE model used for data generation. The predictor matrix $\mathbf{X}$ consisted of six variables generated from a multivariate normal distribution with mean vector zero and covariance matrix:

$$\sigma = \begin{bmatrix} 1 & 0.6 & 0.6 & 0 & 0 & 0 \\ 0.6 & 1 & 0.6 & 0 & 0 & 0 \\ 0.6 & 0.6 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.6 & 0.6 \\ 0 & 0 & 0 & 0.6 & 1 & 0.6 \\ 0 & 0 & 0 & 0.6 & 0.6 & 1 \end{bmatrix}$$

to induce strong correlations among predictors within each block while maintaining independence across blocks. Such a configuration reflects the ERA-GEE component structure, where each predictor contributes to a prespecified component and the resulting components are orthonormal. Component weights and regression coefficients were specified as:

$$\mathbf{W} = \begin{bmatrix} 0.71 & 0 \\ 0.57 & 0 \\ 0.42 & 0 \\ 0 & 0.71 \\ 0 & 0.57 \\ 0 & 0.42 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0.5 & 0.5 & 0.8 \\ 0.9 & 0.7 & 0.6 \end{bmatrix}.$$

The original (pre-normalization) component weights for each predictor block were specified as $w_k = \{0.5, 0.4, 0.3\}$, and were subsequently normalized to unit length to satisfy the ERA

orthonormality constraint. These weights were selected to reflect differential predictor contributions to each component. The regression coefficients were specified such that the resulting outcome covariance structure closely approximates an exchangeable correlation pattern with pairwise correlations of approximately 0.7. This level of correlation was chosen to resemble the magnitude observed among outcomes in the CLSA empirical application used later in this thesis. Component scores were calculated as $F = XW$ and subsequently scaled. Outcomes were then generated according to:

$$Y = FB + \varepsilon, \varepsilon_{ij} \sim N(0, 0.5^2).$$

Missing-Data Design and Implementation

To evaluate each missing-management approach—LWD, MICE, and LSI-ERA-GEE—, missingness was introduced into the simulated datasets under three mechanisms: MCAR, MAR, and MNAR. Missingness was generated using a weighted-sampling procedure that allowed precise control over the proportion of missing values in each variable. Let $\mathbf{Z}$ denote the full data matrix containing all predictors and outcomes, with dimensions N by $(P + Q)$. Let R_{ij} be a missingness indicator defined as

$$R_{ij} = \begin{cases} 0 & \text{if } \mathbf{Z}_{ij} \text{ is missing,} \\ 1 & \text{if } \mathbf{Z}_{ij} \text{ is observed,} \end{cases}$$

and p denote the proportion of missing values to be introduced in each column.

For MCAR, missing values were introduced by randomly sampling $N \times p$ rows from each column without weights. The probability model is therefore:

$$\Pr(R_{ij} = 0) = p.$$

This method generates missingness that is MCAR, because the probability of a value being missing in any given variable is completely independent of any other observed variables, and the unobserved values themselves.

For MAR, the probability of missingness in a variable was determined by the observed value of a different variable. To construct this mechanism, a standardized copy of the initially complete dataset, denoted $\tilde{\mathbf{Z}}^{obs}$, was used to generate sampling weights. The sampling weights were generated with a positive mechanism such that rows with higher values in each column received larger weights. To implement the positive mechanism, softmax weighting was utilized to assign higher values within a standardized variable exponentially larger weights, which were then used as sampling weights. Scaling constants of two and five were used for MAR and MNAR, respectively, to produce moderate dependencies for each mechanism. $N \times p$ rows from each column of $\tilde{\mathbf{Z}}^{obs}$ were sampled using the weights, then the sampled rows in each column were assigned to have missing data on a different column in $\mathbf{Z}$. The probability model under the MAR mechanism is therefore:

$$\Pr(R_{ij} = 0 | \mathbf{Z}_{i-j}^{obs}) = \frac{\exp(2\tilde{\mathbf{Z}}_{i-j}^{obs})}{\sum_{i=1}^n \exp(2\tilde{\mathbf{Z}}_{i-j}^{obs})}.$$

This creates missingness that is MAR, as the probability of missingness on a given variable is completely dependent on the observed value of a different variable.

For MNAR, the probability of missingness in a variable was determined by the initially observed value from the same variable before missingness was introduced. Sampling weights for each variable were generated using their own observed values with a positive mechanism. Again, $N \times p$ rows from each variable were sampled using the weights, and the sampled rows were assigned to have missing data. The probability model under the MNAR mechanism is therefore:

$$\Pr(R_{ij} = 0 | \mathbf{Z}_{ij}) = \frac{\exp(5\tilde{\mathbf{Z}}_{ij})}{\sum_{i=1}^n \exp(5\tilde{\mathbf{Z}}_{ij})}.$$

Once the missingness has been introduced, the probability of missingness is now entirely dependent on the unobserved values, thereby satisfying the definition of MNAR.

Experimental Conditions

The simulation study followed a full factorial design involving four experimental factors:

1. Missing mechanism: MCAR, MAR, MNAR
2. Proportion of missing data: 5%, 10%, 20%, 30%
3. Sample size: $N \in \{50, 100, 300, 500\}$
4. Missing-management method: LWD, MICE, LSI-ERA-GEE

Each simulation factor pertaining to the missing data was selected to stress-test each missing-management method under realistic missing-data conditions, and the levels were chosen based on prior simulation work that used comparable scenarios (e.g., Choi et al., 2020; Dittlhong et al., 2018; Kim & Hwang, 2022; Wasito & Mirkin, 2006). Missing mechanism was selected as a factor to evaluate the sensitivity of results to various mechanisms, proportion of missing data was manipulated to assess robustness of the results to increasing levels of information loss, and sample size was varied to evaluate the algorithm's stability and estimation accuracy with different amounts of information available. Lastly, missing-management method was varied to directly compare how each method influences estimator bias, efficiency, and convergence.

Crossing mechanism $\times$ proportion $\times$ sample size $\times$ management method yielded $3 \times 4 \times 4 \times 3 = 144$ unique conditions. For each condition, 1,000 Monte Carlo iterations were performed. Within each iteration, a complete dataset was first generated according to the ERA-GEE model described above. Missing values were then introduced according to the specified mechanism and

proportion, after which the selected missing-management method was applied. Finally, the ERA-GEE model was fit, and the resulting parameter estimates and diagnostic outputs were stored for analysis. When fitting the ERA-GEE model, regularization was not applied—both tuning parameters were set to zero. This was done so the missing data conditions and missing-management methods would be the only factors influencing the estimated component weights and regression coefficients.

All conditions and iterations of the simulation were completed in R (version 4.5.2) using parallel computation on a Mac Mini desktop computer with 10 processing cores. Nine cores were used to execute simulation conditions in parallel, while one core was reserved for system processes. Each core processed one experimental condition at a time, completing all 1,000 iterations before proceeding to the next condition. For reproducibility, all R code used to run the simulation including the required custom functions can be found on the GitHub page linked here <https://github.com/QuantMM/Missing-Management-in-ERA-GEE/tree/main>.

Performance Metrics

The simulation results were evaluated using two groups of performance metrics: feasibility metrics which assessed the stability and computational behaviour of each missing-management method within ERA-GEE estimation, and estimation quality metrics, which evaluated the statistical properties of the resulting parameter estimates. To assess the feasibility of each missing-management method under different missing-data conditions, three metrics were examined: convergence probability, convergence time, and iterations to convergence.

The convergence probability was defined as the proportion of Monte Carlo iterations in which the ERA-GEE algorithm successfully converged and returned estimates for all model parameters. For each experimental condition, convergence probability was calculated as the

number of converged iterations divided by the total number of iterations (1,000). Values closer to one indicate greater stability of the estimation procedure under the given missing-data scenario.

The convergence time was defined as the mean computational time required for the algorithm to reach convergence across iterations within a condition. This metric provides information on the relative computational efficiency of each missing management method, where smaller values indicate faster estimation.

The number of iterations in the ERA-GEE ALS algorithm to convergence was recorded for each Monte Carlo iteration and averaged across the 1,000 iterations in each condition. This metric reflects the convergence efficiency of the estimation procedure. To contextualize the iteration counts, the average total-loss value at convergence—defined as the sum of squared errors from the ERA-GEE objective function—was also recorded. This metric provides insight into the level of objective-function minimization achieved when convergence occurred.

The statistical quality of the parameter estimates was evaluated using variance, bias, and root mean squared errors (RMSE). Variance measures the dispersion of parameter estimates across Monte Carlo iterations and reflects estimator precision. The variance of parameter j was calculated as:

$$Var = \frac{1}{I-1} \sum_{i=1}^I (\hat{\theta}_{ij} - \bar{\theta}_j)^2,$$

where $\hat{\theta}_{ij}$ denotes the estimate of parameter j in iteration i , $\bar{\theta}_j$ denotes the mean estimate across all I iterations, and I is the total number of iterations. To simplify interpretation, variance values were pooled across parameters within each parameter group. Specifically, variances were averaged across all regression coefficients and across all component weights within each iteration. These pooled values were then averaged across the 1,000

iterations, yielding a single variance measure for regression coefficients and component weights in each condition.

Bias measures the average deviation of an estimator from the true parameter value used in data generation. The bias for parameter j was calculated as:

$$Bias = \frac{1}{I} \sum_{i=1}^I (\hat{\theta}_{ij} - \theta_j),$$

where θ_j is the true value of parameter j . Bias provides information on the accuracy of an estimator, where values closer to zero indicate greater accuracy. Similar to variance, bias values were pooled across regression coefficients and component weights within each iteration and then averaged across the 1,000 iterations.

RMSE combines information about both estimator bias and variability. The RMSE for parameter j was calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^I (\hat{\theta}_j - \theta_j)^2}{i}}.$$

Theoretically, RMSE can be expressed as the square root of the sum of squared bias and variance, i.e., $RMSE = \sqrt{Bias^2 + Var}$. Bias and variance typically have an inverse relationship, therefore, RMSE is an important hybrid performance metric that aims to capture the bias-variance trade-off. As such, RMSE provides joint information on both the precision and accuracy of parameter estimates, with lower values indicating better performance. RMSE values were also pooled for regression coefficients and component weights within each iteration and then averaged across the 1,000 iterations.

Empirical Application

Data

The empirical application used data from the CLSA which was approved by the research ethics boards of all participating institutions. The present analysis also received approval from the University of Manitoba Research Ethics Board (Protocol Number: HE2026-0037). The analytic sample included 30,097 individuals from the comprehensive cohort observed at the baseline, follow-up 1, and follow-up 2 timepoints. Because the primary goal of the empirical application was to evaluate the effectiveness of different missing-management methods, all individuals with missing data were included in the analysis, and no exclusion criteria were used.

Variables

The baseline predictor variables used in the analysis were income, education, savings, diet, strength training, walking, social outings, volunteering, and loneliness. The outcome variables included in the analysis were executive function scores measured at baseline, follow-up 1, and follow-up 2. In the CLSA, executive function scores are calculated as composite scores derived from four cognitive tasks assessing executive functioning. Detailed descriptions of these tasks, their scoring procedures, and the construction of the executive function summary scores are provided in O’Connell et al. (2022). Measurement scales and descriptive statistics for all variables are presented in Table 1.

Model Specification

The ERA-GEE model used in the empirical analysis consisted of three predictors per component, three components, and three outcomes. Predictors were grouped into three conceptual domains—SES, health behaviours, and quality of social life—based on prior research

indicating that cognitive aging is influenced by multiple domain-specific components (e.g., Meister & Zahodne, 2022; Stinchcombe et al., 2023; Zaninotto et al., 2018). Component scores were estimated from the predictor blocks and used to predict the longitudinal outcomes within the ERA-GEE framework. With this empirical application revolving around domain-specific components and repeated measures of a longitudinal outcome, this analysis is well-suited for highlighting ERA-GEE's applicability to real-world data. Additionally, the selected variables all exhibit significant within-block correlations, allowing for examination of how each missing management method handles correlated high-dimensional predictors (within-block correlations are presented in Table 2).

Three separate identical ERA-GEE analyses were conducted, each with a different missing-management method applied to allow for the direct comparison of how each method impacts the analysis. In the analyses that used LWD and MICE, missing values were managed before the ERA-GEE model was fit, while missing values in the LSI-ERA-GEE analysis were managed within the model fitting procedure. As in the simulation, both regularization parameters were set to zero so the missing-management method would be the only factor influencing the results across analyses. For statistical inference, bootstrapping with 1,000 replications where individuals were resampled with replacement to preserve the within-subject dependence structure was conducted using each "complete" dataset following the deletion/imputation of missing values.

Missing-Data Exploration

An exploratory missing-data analysis was conducted to examine the extent of the missingness in the dataset prior to any missing management and gain insights into what missing mechanisms could reasonably be inferred. The exploratory analysis included examinations of the

proportion of missing values across each variable individually and across the entire dataset, and using logistic missingness models to support inferences about the underlying missing mechanism.

The proportion of missing values in each variable was calculated as the number of NA values in a variable divided by the number of observations in the sample ($N = 30,097$). This proportion provides information on how much data is missing in each variable and tells us which variables are most impacted by missingness. The proportion of missing values across the entire dataset was calculated as the total number of NA values divided by the total number of values in the entire dataset ($N \times (P + Q)$). This value tells us how much of the entire dataset is missing, and allows us to examine the severity of missingness before and after LWD is applied.

Logistic missingness models were used to provide data-driven insights into what missing mechanism may underlie the missing data. Although it is not possible to definitively prove which missing mechanism underlies the missingness due to the data's unobserved nature, logistic regression models with binary missingness indicators as outcomes can be used to obtain evidence in support of the assumption that the missingness is MAR (Fairclough, 2012, Chapter 6). In these models, observed variables are used as predictors of missingness in other variables via logistic regression. If any predictors have slope coefficients that are statistically significant, this indicates that missingness in the current outcome variable is associated with other observed variables, which satisfies the definition of MAR. Therefore, we can reasonably infer that significant predictors in these models provide evidence that the underlying mechanism is MAR.

In the empirical analysis, five variables with the largest proportions of missing values were selected to be examined using logistic missingness models. In each model, a binary missingness indicator variable was created based on the variable with missing data, and a logistic

regression model was fit using all other variables as predictors. Results from these models were used to support inferences about the underlying missing mechanism in the dataset.

Evaluation of Missing-Management Methods

To evaluate the impact each missing-management method had on the results of the ERA-GEE analysis, four key metrics were compared across each analysis: estimator standard errors, statistical significance of estimators, confidence interval overlap, and predictive performance. To obtain information on the precision of parameter estimates, the standard error of each estimator was calculated as the standard deviation of each estimator's bootstrap distribution. The standard error tells us how much we can expect each estimator to vary across random samples from the population, with smaller values indicating greater precision.

To examine if the results of the analysis change based on the missing-management method used, the statistical significance of each estimator was compared across all analyses. The statistical significance of each estimator was determined by constructing bootstrapped confidence intervals using the 5th and 95th percentiles. Confidence intervals containing zero indicate the estimator is not statistically significant, while intervals not containing zero indicate the estimator is significant. The degree of dis/agreement in the significance of estimators across the analyses tells us if the results depend on the missing-management method used.

To determine how much the results change depending on the missing-management method used, the overlap of confidence intervals for each estimator was also compared across the three analyses. Evaluating overlap in the confidence intervals provides information on if the magnitude of estimators differs based on the missing-management method used. A complete lack of overlap between confidence intervals indicates the magnitude of estimators are considerably different.

The RMSE of model-predicted versus observed outcomes was calculated to examine predictive performance. The RMSE was calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\hat{Y} - Y)^2}{N}},$$

where $\hat{Y}$ is an N by Q matrix of model-predicted outcomes, and Y is the matrix of observed outcomes. This RMSE value tells us how well the model predicts the outcomes given the final parameter values at convergence, with lower values indicating more accurate prediction.

Results

Simulation Results

Method Feasibility

The feasibility of each missing-management method was evaluated by examining convergence probabilities, convergence time, number of iterations to convergence, and the total-loss value at convergence across all experimental conditions. All results pertaining to the feasibility metrics from the simulation are presented as tables in Appendix B. As shown in Table B1, overall, both MICE and the proposed LSI-ERA-GEE approach demonstrated stable convergence behaviour across all simulation scenarios, whereas LWD frequently failed to converge when missingness was high and sample sizes were small.

Across all missing mechanisms, MICE and LSI-ERA-GEE achieved perfect convergence probabilities in every condition. In contrast, LWD exhibited reduced convergence probabilities under MCAR and MAR when sample sizes were small and the proportion of missing values was high. Examination of the retained sample sizes shown in Table B2 revealed that in these conditions, fewer than 15% of the observations remained after deletion, indicating the

convergence failures under LWD were primarily driven by severe reductions in effective sample size. These results indicate that MICE and LSI-ERA-GEE always yield estimable solutions, even under intensive missing-data conditions where LWD struggles.

Table B3 shows the mean convergence time in each condition. As the proportion of missing values increased, convergence times for both MICE and LSI-ERA-GEE also increased, reflecting the additional computational effort required to handle missing data. In contrast, convergence time for LWD decreased as missingness increased because fewer observations remained after deletion. Overall, LWD converged substantially faster than both MICE and LSI-ERA-GEE because it reduces the size of the dataset prior to estimation. Between the two imputation-based methods, LSI-ERA-GEE generally required less computation time than MICE under low-to-moderate missingness, whereas MICE was slightly faster under the highest levels of missingness.

As shown in Table B4, the number of iterations required for convergence increased as the proportion of missing data increased for all methods, although this pattern was most pronounced for LSI-ERA-GEE. Importantly, despite requiring more iterations, LSI-ERA-GEE consistently achieved the lowest total-loss values at convergence (total-loss values at convergence are presented in Table B5). This indicates that the additional iterations correspond to improved optimization of the ERA-GEE objective function, rather than computational inefficiency. Taken together, these feasibility results indicate that LSI-ERA-GEE provides stable convergence behaviour across a wide range of missing-data conditions while maintaining reasonable computational cost.

Estimator Behaviour

Estimator behaviour was evaluated using variance, bias, and RMSE across all simulation conditions. All results pertaining to the estimation quality metrics from the simulation are presented as tables in Appendix C. Variance was analyzed to assess the precision of parameter estimates. The variance results for coefficient and weight estimates are summarized in Tables C1–3. Across all three missing-management methods and missing mechanisms, two dominant structural patterns emerged. First, estimator variance was strongly influenced by sample size. Across all missing mechanisms and estimation methods, variance decreased substantially as the sample size increased. This pattern indicates that larger samples consistently produce more stable parameter estimates regardless of the missing-management method used. Second, variance increased systematically as the proportion of missing values increased. This pattern was evident across all missing mechanisms, indicating that information loss due to missingness plays a primary role in inflating estimator variability.

Although these general patterns were consistent across methods, the magnitude of the variance differed substantially between missing-management methods. The MICE estimator maintained relatively low and stable variance across missing-data conditions, increasing only modestly as the proportion of missing data grew. In contrast, the LWD estimator showed substantially greater sensitivity to missingness: while variance was comparable to other methods at low levels of missing data, it increased sharply as missingness rose, particularly in smaller samples and when missingness exceeded 20%. The LSI-ERA-GEE estimator also exhibited increasing variance as missingness rose, although the magnitude of this increase was generally smaller than that observed for LWD. Overall, these results indicate that estimator variance is primarily driven by two factors: increasing sample size substantially stabilizes parameter

estimates, whereas increasing proportions of missing data inflate estimator variability, with the LWD estimator showing the greatest sensitivity to missingness.

The bias results for coefficient and weight estimates are summarized in Tables C4–6. In contrast to the variance results, bias appeared largely insensitive to sample size. Across all sample sizes, the overall bias patterns remained nearly identical, although the magnitude of bias tended to be somewhat larger in smaller samples. Instead, bias was primarily influenced by the proportion of missing data and the underlying missing mechanism.

When comparing the estimators, the MICE estimator exhibited relatively stable bias across most conditions. Under both MCAR and MAR mechanisms, the bias remained small and approximately constant across increasing levels of missingness, indicating that MICE produced estimates close to the true parameter values even as missingness increased. In contrast, the LSI-ERA-GEE estimator displayed a clear sensitivity to missingness. While bias remained small at low-to-moderate levels of missing data, it shifted progressively in the positive direction as the proportion of missing values increased, indicating increasing upward bias in the estimated parameters.

The largest bias values were observed for the LWD estimator under conditions with high missingness and small sample sizes. When the proportion of missing data exceeded 20%, LWD produced substantial positive bias, particularly when the sample size was small (e.g., $n = 50$ or $n = 100$). This instability reflects the severe reduction in effective sample size resulting from LWD.

Under the MNAR mechanism, the divergence between estimators became more pronounced. The MICE estimator exhibited increasingly negative bias as missingness increased, indicating systematic underestimation when the missing mechanism depended on unobserved

values. In contrast, the LSI-ERA-GEE estimator showed the opposite pattern, with bias increasing in the positive direction as missingness grew. These results suggest that violations of the MAR assumption can affect estimators in fundamentally different ways.

Finally, RMSE was analyzed to evaluate the overall quality of parameter estimates by jointly accounting for both variance and bias. Tables C7–9 provide the RMSE values for both regression coefficients and component weights in all conditions. Across all simulation conditions, RMSE increased as the proportion of missing values increased and decreased as sample size increased, reflecting the combined influence of information loss and estimator stability. Despite larger variance observed for LSI-ERA-GEE in some scenarios, its relatively low bias allowed it to achieve competitive or lower RMSE values in many conditions.

In almost all scenarios with small-to-moderate proportions of missing values, LSI-ERA-GEE exhibited the lowest RMSE for both regression coefficients and component weights. However, in a small number of conditions characterized by small sample sizes and moderate missingness, MICE achieved slightly lower RMSE values than LSI-ERA-GEE. When the proportion of missing values exceeded 20%, MICE tended to produce the lowest RMSE for regression coefficients, whereas LSI-ERA-GEE continued to exhibit the lowest RMSE for component weights across all conditions.

Importantly, LWD never produced the lowest RMSE in any simulation condition, indicating that deletion-based estimation consistently resulted in poorer overall estimator performance relative to imputation-based missing-management methods. Taken together, these results demonstrate that while the estimators exhibit different variance and bias patterns, the LSI-ERA-GEE procedure provides highly competitive overall estimation performance and is

particularly effective at recovering component weights even under complex missing-data scenarios.

Empirical Application Results

Missing-Data Analysis

Prior to conducting the empirical analysis, the overall extent and structure of missingness in the dataset were examined. The proportion of missing values for each variable is presented in Table 1. The largest proportions of missingness occurred in the cognition outcomes measured at the three timepoints, as well as in the baseline education and savings variables. Missingness in these variables exceeded 15%, whereas all remaining variables exhibited relatively low levels of missingness, generally below 6%. Across the entire dataset, 13.22% of all values were missing. Importantly, this 13.22% of missing values was distributed across 23,954 individuals, indicating only 6,143 individuals had complete data. Consequently, under LWD, approximately 80% of the original sample was deleted from the analysis.

To investigate the underlying missing mechanism, logistic missingness indicator models were estimated for the five variables with missingness exceeding 15%. For cognition at follow-up 1 and follow-up 2, as well as education and savings, the missingness indicators were significantly associated with at least two observed predictors. These results suggest that the probability of missingness in these variables depends on observed information, providing evidence that the missing mechanism is at least partially consistent with the MAR assumption.

In contrast, missingness in baseline cognition was not significantly associated with any of the observed variables, indicating the missingness was either completely random, or associated with the unobserved values themselves. However, due to the unobserved nature of the missing values, it is not possible to directly test whether the missingness depends on the unobserved

values. Overall, the results indicate that most missingness in the dataset is systematically related to observed variables, making the MAR assumption plausible, although the possibility of MNAR cannot be entirely ruled out.

Bootstrap Convergence Behaviour

To examine the computational behaviour of each missing-management method in a realistic setting, the bootstrap estimation procedure was implemented for each method, and two key computational characteristics were recorded. Specifically, the mean total-loss value at convergence and the total computation time of the bootstrapping procedure were compared across methods.

The mean total-loss value at convergence was 2.78 for LWD, 2.76 for MICE, and 1.66 for LSI-ERA-GEE, indicating the LSI-based approach converged to substantially lower objective function values than the other methods. This result suggests that the LSI-based estimation achieved a more optimal fit of the ERA-GEE objective function.

In terms of computational efficiency, LWD required the shortest computation time, with the entire bootstrapping procedure being completed in 0.57 hours. MICE required substantially more computation time with the bootstrapping procedure taking 16.83 hours to complete. LSI-ERA-GEE required 1.82 hours in total, representing a moderate computational cost that was substantially lower than MICE. These results indicate that although deletion-based estimation is computationally fastest, imputation-based methods provide more stable optimization behaviour, with LSI-ERA-GEE achieving the most favourable total-loss values while maintaining reasonable computational cost.

Estimator Comparison

To provide context for the empirical results, Figure 2 demonstrates the distribution of regression coefficient bias from the simulation study under conditions most similar to the empirical dataset, namely large sample size, moderate levels of missingness, and mechanisms that are not completely random. This comparison provides a benchmark for evaluating the expected performance of each missing-management method when the true underlying missing mechanism is known. Across both missing mechanisms examined in Figure 2 (MAR and MNAR), LSI-ERA-GEE consistently exhibited the lowest bias in the estimation of regression coefficients. While the simulation study allows direct evaluation of estimator accuracy due to known true parameter values, the empirical application focuses on comparative properties such as stability and predictive performance, as the true data-generating mechanism is unknown. Therefore, these results suggest that when the missingness resembles the empirical setting, LSI-ERA-GEE is expected to produce the best predictive performance among the methods considered.

Table 3 includes the parameter estimates, confidence intervals, and standard errors for all component weights and regression coefficients from each missing-management method. For the component weight estimates, LSI-ERA-GEE consistently produced the smallest standard errors relative to LWD and MICE, indicating LSI-ERA-GEE provided the most precise estimates of the component structure. In contrast, LWD exhibited the largest standard errors across all component weights, reflecting the loss of information resulting from deletion of incomplete observations.

Under the LSI-ERA-GEE analysis, all component weights were statistically significant, as the 95% confidence intervals did not include zero. In contrast, some component weights were not statistically significant in the MICE and LWD analyses. This difference strongly suggests

that the choice of missing management method can substantially influence inference regarding the component structure within an ERA analysis.

Even in cases where component weights were statistically significant in all three analyses, the magnitude of the estimated weights differed considerably. For example, the 95% confidence intervals for the weights of income and savings obtained from the LWD and MICE analyses did not overlap with the corresponding confidence intervals obtained from the LSI-ERA-GEE analysis. Within ERA-GEE, all variables and components are standardized using a common procedure making all parameter estimates directly comparable across methods in terms of scale and interpretation—estimated coefficients represent quantities on a common scale. The lack of overlap in confidence intervals therefore indicates noticeable differences in the magnitude of the estimated weights for income and savings within the SES component. A visual representation of this difference in magnitudes is provided in Figure 3.

A similar pattern emerged for the regression coefficients linking the components to the cognition outcomes. LSI-ERA-GEE produced the smallest standard errors for most regression coefficients, whereas LWD again produced the largest standard errors. Although all regression coefficients were statistically significant across all three analyses, there were many instances where confidence intervals for common coefficients did not overlap. While the confidence intervals for estimates from the LWD and MICE analyses overlapped with each other, there was no overlap between the confidence intervals obtained from LWD and LSI-ERA-GEE, and only a single instance of overlap between MICE and LSI-ERA-GEE. This indicates that the magnitude of estimated coefficients across analyses were noticeably different. Figure 4 illustrates this pattern using the estimated effect of SES on cognition at follow-up 2 as an example.

Prediction Accuracy

RMSE was calculated to compare the outcome prediction accuracy in each analysis. LSI-ERA-GEE had the lowest prediction error, with an RMSE of 0.004, followed by MICE with an RMSE of 0.005, and LWD with the highest RMSE of 0.012. These results indicate that the parameter estimates obtained from the LSI-ERA-GEE analysis resulted in the best predictive performance regarding model-based predictions of the cognition outcomes.

Substantive Results from the LSI-ERA-GEE Model

Figure 5 demonstrates a schematic representation of the ERA-GEE model estimated using the LSI-ERA-GEE method and displays all parameter estimates obtained from the empirical analysis. The component structure shows that each predictor contributes significantly to its corresponding component, although each component is primarily dominated by a single key predictor. Education exhibited the largest weight within the SES component, diet contributed most strongly to the health behaviours component, and loneliness had the largest weight within the social life component.

All three components—SES, health behaviours, and social life—exhibited statistically significant positive regression coefficients for all three cognition outcomes, indicating that higher SES, healthier behavioural patterns, and higher quality social engagement are associated with improved cognitive functioning over time. Among the three components, SES consistently exhibited the largest regression coefficients across all timepoints. This pattern suggests that SES plays the most influential role in predicting cognitive performance. Moreover, the magnitude of the SES regression coefficients increased across the three cognition measurements, indicating that the influence of socioeconomic conditions on cognitive outcomes becomes progressively stronger over time.

Overall, the empirical application demonstrates that the choice of missing-management method can substantially influence both statistical inference and substantive interpretation within the ERA-GEE framework. While the conventional deletion-based missing-management method produced the least stable results and largest prediction error, both imputation-based approaches provided more reliable estimates. In particular, the LSI-ERA-GEE procedure consistently yielded more precise parameter estimates, stronger and more stable component structures, and the highest predictive accuracy. These findings are consistent with the simulation results and further illustrate that LSI-ERA-GEE offers a promising integrated approach for managing missing data in high-dimensional longitudinal applications, particularly when missingness is low to moderate, though caution is warranted as the proportion of missingness increases.

Discussion

The overarching purpose of this thesis was to examine and extend missing-data management in the context of ERA-GEE analyses involving high-dimensional longitudinal data. Specifically, this work pursued three core objectives: (1) to develop a novel extension of the ERA-GEE framework that internally manages missing data, (2) to evaluate its performance relative to existing missing-management methods under realistic missing-data scenarios, and (3) to demonstrate the practical implications of these approaches through an empirical application.

The first objective was achieved by extending ERA-GEE's ALS algorithm to include the estimation of missing predictors and outcomes via LSI. Results from both the simulation study and empirical application indicate that the proposed method is not only operationally feasible, but also demonstrates strong and stable performance. The second objective was achieved by conducting a comprehensive Monte Carlo simulation study, in which key missing-data conditions were systematically manipulated. This design enabled a detailed comparison of

competing methods across a range of realistic scenarios, yielding insights into their respective strengths and limitations. The third objective was achieved by conducting an empirical analysis of cognition data from the CLSA, applying three different missing-management strategies within the ERA-GEE framework. This application provided a direct illustration of how analytical results can vary depending on the choice of missing-management method. Taken together, the findings from both the simulation study and empirical analysis provide a comprehensive evaluation of missing-management strategies in ERA-GEE, advancing methodological understanding of their performance in high-dimensional longitudinal settings and informing their application in empirical research.

LWD

The results from the simulation and empirical application both provide evidence that LWD is a suboptimal missing-management method for ERA-GEE analyses. Despite LWD's low computational burden and fast convergence times, these advantages are outweighed by substantial reductions in retained sample size. In the simulation, even with the smallest amount of missing data (5%), applying LWD resulted in losing between 21–37% of the sample. This issue was even more pronounced in the empirical analysis. Although 13% of the original data was missing, applying LWD reduced the retained sample by approximately 80%. This pattern suggests that even with moderate levels of missing data, applying LWD can translate into substantial case loss in high-dimensional settings. This is because the more variables there are in an analysis, the more opportunities there are for missingness to occur within an individual's record, thereby decreasing the probability that any given individual has complete data (J. Wang & Aronow, 2023).

Two key consequences of such large reductions in sample sizes are inflated variance of estimates and decreased representativeness of the analytic sample. The variance inflation of estimates was evident in both the simulation and empirical analysis. In nearly all conditions of the simulation, LWD exhibited the highest variance of parameter estimates compared to MICE and LSI-ERA-GEE. Similarly, in the empirical analysis, LWD exhibited the largest standard errors and widest confidence intervals across all parameters. This pattern is consistent with prior literature showing that LWD inflates the variability of estimates through sample-size reduction (e.g., Huque et al., 2018; Ji et al., 2018; Steif et al., 2022). Smaller retained sample sizes following deletion also cause complications regarding statistical inference. When the goal of an analysis is to draw conclusions about a population of interest, it becomes difficult to assume that the analytic sample remains representative after a substantial proportion of the original sample has been removed (Feng et al., 2012; Woods et al., 2024).

Taken together, these findings indicate that LWD is not well-suited for ERA-GEE analyses in high-dimensional longitudinal settings. Rather than preserving statistical validity, LWD exacerbates data loss and amplifies variability in parameter estimates. Accordingly, it should not be relied upon as a default missing-management strategy in this context.

MICE

A consistent pattern in the simulation was that MICE exhibited the lowest variance across nearly all conditions. Likewise, MICE also maintained low standard errors in the empirical analysis. This stability is consistent with prior literature suggesting that, relative to other missing-management methods, MICE produces more precise estimates (Azur et al., 2011; Feng et al., 2012). One key reason for this performance lies in its treatment of uncertainty. Rather than imputing each missing value once and deterministically treating it as observed, MICE generates

multiple estimates that each contribute to parameter estimation via the pooling procedure. This approach incorporates imputation uncertainty into the final estimates (Azur et al., 2011; van Buuren & Groothuis-Oudshoorn, 2011), while reducing the influence of extreme estimates across imputations and stabilizing the resulting parameter estimates.

Additionally, MICE also exhibited a favourable bias-variance trade-off in the simulation and a competitive level of prediction error in the empirical application. More specifically, in a direct comparison of RMSE, MICE consistently outperformed LWD across all simulation conditions. A similar pattern emerged in the empirical analysis, where MICE yielded lower prediction error than LWD. These findings indicate that MICE provides both more accurate and stable estimates than LWD, which is consistent with observations from prior research (e.g., Huque et al., 2018; Ji et al., 2018; Gad, 2017).

Taken together, these results suggest that although MICE is not fully aligned with the structural assumptions of ERA-GEE, it remains a more effective missing-management strategy than LWD in high-dimensional longitudinal settings. While its imputation models do not explicitly account for component structure, MICE better preserves sample information and produces more stable and accurate estimates than LWD. Accordingly, MICE represents a viable baseline approach for ERA-GEE analyses and should be preferred over LWD when more tailored methods are not available.

LSI-ERA-GEE

Across both the simulation and empirical application, LSI-ERA-GEE demonstrated stable convergence and reasonable computation times. As this novel method extended ERA-GEE's ALS algorithm, it was initially unclear whether the estimation procedure would converge reliably, particularly under high levels of missingness. As the proportion of missing data

increases, larger portions of the dataset are updated iteratively, which could potentially hinder convergence. Despite this concern, LSI-ERA-GEE exhibited consistent and stable convergence across a wide range of missing-data conditions, indicating that the estimation procedure is both viable and stable. Although the method required longer computation times than LWD and, in some cases, MICE, these differences were modest, suggesting that it remains computationally feasible even under demanding settings.

In the empirical application, where statistical inference was required, standard errors and confidence intervals were obtained via bootstrapping. Within this context, a substantial computational difference emerged across methods. In particular, the bootstrap procedure required considerably more time for MICE than for LSI-ERA-GEE. Because MICE generates multiple imputed datasets, the full bootstrap procedure must be repeated separately for each dataset, leading to a multiplicative increase in computational burden. As a result, the total computation time for MICE was approximately nine times longer, corresponding to a difference of nearly 15 hours in the empirical analysis. In contrast, LSI-ERA-GEE integrates the imputation and estimation steps within a single iterative procedure, allowing the bootstrap to be applied directly without repetition across imputed datasets. This integration not only reduces computation time, but also simplifies implementation. In the MICE framework, researchers must specify imputation models, generate multiple imputed datasets, fit the analysis model to each dataset, and pool the results. In comparison, LSI-ERA-GEE provides a more streamlined workflow, making it both computationally efficient and easier to implement in practice.

Beyond its stable convergence and computational feasibility, LSI-ERA-GEE demonstrates strong performance in handling missing data, particularly under low-to-moderate levels of missingness. In the simulation study, LSI-ERA-GEE produced variances comparable to

MICE while achieving the lowest bias among the methods considered in low-missingness conditions. Consistent with these findings, the empirical application showed that LSI-ERA-GEE yielded the smallest standard errors and prediction error relative to LWD and MICE.

However, a key limitation of LSI-ERA-GEE is that its performance severely degrades under higher proportions of missing data. As the proportion of missing values reaches 20% or higher, LSI-ERA-GEE exhibits severe bias and variance inflation. This pattern is likely explained by the fact that LSI-ERA-GEE is a single-imputation method, as the literature indicates that deterministic imputation methods can inflate the variance and bias of parameter estimates (Feng et al., 2012). Although LSI-ERA-GEE uses an estimation procedure that iteratively updates missing values until convergence, the imputed values at convergence are treated as fixed and observed, without any consideration of the uncertainty due to the missing data.

This limitation can be understood by examining the distribution of parameter estimates across simulation iterations. Under LSI-ERA-GEE, poor imputations directly translate into extreme parameter estimates which remain in the distribution, and effectively act as outliers inflating both bias and variance. In contrast, MICE mitigates this effect by averaging estimates across imputed datasets, thereby reducing the influence of such extremes. Consequently, the observed variance inflation under LSI-ERA-GEE at higher levels of missingness reflects a fundamental limitation of its single-imputation framework.

Overall, the empirical application and simulation both provide evidence that LSI-ERA-GEE maintains solid convergence behaviour, reasonable computation time, and effective missing-data management in certain conditions. More specifically, among the methods considered in the present study, LSI-ERA-GEE exhibited the best performance under small-to-

moderate proportions of missing values. By integrating LSI within the ERA-GEE framework, this extension removes the requirement of complete data for estimation and inference, substantially improving the practical applicability of ERA-GEE in real-world settings. At the same time, its limitations under higher levels of missingness should be carefully considered, as they reflect the inherent limitations of single-imputation approaches. Accordingly, LSI-ERA-GEE is best suited for settings in which missingness is not excessive, while methods such as MICE may be preferable when the proportion of missing data becomes substantial (e.g., $\geq 20\%$).

Although LSI-ERA-GEE's performance appears to be most sensitive to the proportion of missing values, the underlying missing mechanism is also important to consider. Under low-to-moderate proportions of missing values, LSI-ERA-GEE remains valid under both MCAR and MAR mechanisms, however, when the missingness is non-random, the model's estimates exhibit considerable positive bias as the proportion of missing values increases. This indicates that LSI-ERA-GEE is not robust to violations of the MAR assumption. Therefore, if researchers have reason to believe that missingness in their data is non-random, they should include auxiliary variables that are strongly correlated with the variables containing missingness to strengthen the MAR assumption (Collins et al., 2001), or use missing-management methods that were specifically developed for handling non-random missingness such as selection or pattern-mixture models. In such scenarios, if the goal of the analysis is to make inferences about a population of interest, conclusions should be made with a high level of caution and transparency regarding the missing data and its management, as non-random missingness has high potential for introducing bias.

Lastly, LSI-ERA-GEE's performance is also influenced by sample size, making this another factor to consider when using this method. LSI-ERA-GEE maintains strong performance

under small sample sizes, however, performance continues to improve as the sample size increases. Overall, under the conditions examined in this thesis, LSI-ERA-GEE should be the preferred method for missing management in ERA-GEE analyses when the proportion of missing data is below 20% and the missing mechanism is believed to be MCAR or MAR; larger sample sizes are preferable but not essential.

Limitations

One limitation of the present study is that the simulation used a relatively simple data-generation process where all variables were normally distributed, the relationships between predictors and outcomes were linear, and no interactions were included. The only complexity came from the dependency between certain predictors due to the within-block correlations. Therefore, the data-generating process in the simulation constitutes a somewhat idealized situation in comparison to applied settings where the true data-generating process is unknown, and relationships among variables may deviate from standard modeling assumptions. Such deviations are important to consider as they can impact the quality of imputed values. For example, when the data-generating process includes nonlinearities, interactions, or higher-order terms, conditional imputation models within MICE become misspecified, leading to poor imputations (Allison, 2002; Seaman et al., 2012). In general, it has been argued that simulation studies should reflect realistic departures from ideal model assumptions to better reflect applied conditions (Burton et al., 2006). This allows for a more informative assessment of how different methods perform in practice, rather than under idealized scenarios. To address this limitation, future work could extend the simulation design to include a more complex data-generating process to evaluate the robustness of all missing-management methods under more realistic conditions.

A second limitation of this study is that only continuous predictors and outcomes were considered in both the simulation and empirical analysis. As a result, the performance of the missing-management methods under mixed or non-continuous variable types remains unclear. Although both MICE and LSI-ERA-GEE can, in principle, accommodate variables from different distributions, this capability was not examined in the present study. Future research could address this limitation by incorporating a broader range of variable types within the ERA-GEE framework to evaluate method performance under more diverse data conditions.

Future Directions

The most important future direction in this line of work is to address the bias and variance inflation observed for LSI-ERA-GEE under higher proportions of missing data. One promising extension is to integrate LSI-ERA-GEE with an MI framework. A potential implementation would involve generating multiple imputed datasets using LSI-ERA-GEE. Instead of relying on deterministic initialization (e.g., column means), stochastic initialization strategies could be used to produce variation in imputed values across runs, as is done in the MICE algorithm (van Buuren & Groothuis-Oudshoorn, 2011). The LSI-ERA-GEE model could then be fit repeatedly to the same incomplete dataset, generating multiple sets of parameter estimates, which would subsequently be combined using Rubin's rules (Rubin, 1987). This approach would allow variability due to missing data to be reflected in the final estimates, rather than relying on a single complete dataset. Future research is needed to evaluate whether such an approach can mitigate the instability observed under higher levels of missingness.

Another important direction involves exploring machine learning (ML)-based methods for missing management. ML methods have emerged in the missing-management literature due to their ability to capture complex and nonlinear relationships. Techniques such as support vector

machines, random forests, and neural networks have been adapted for imputation tasks (e.g., Bertsimas et al., 2018; He et al., 2025; Li et al., 2024). In particular, Generative Adversarial Imputation Nets (GAIN) have been found to outperform MICE, and maintain solid performance under high proportions of missing data (Yoon et al., 2018). These findings suggest that GAIN may have advantages over both MICE and LSI-ERA-GEE. Moreover, prior work has demonstrated that GAIN can be applied within structured modeling frameworks such as structural equation modeling (He et al., 2025), suggesting potential compatibility with ERA-GEE. Given these developments, future research should investigate the integration of ML-based imputation methods within the ERA-GEE framework and evaluate their performance relative to existing approaches under a range of missing-data conditions.

References

- Allison, P. (2002). *Missing data*. SAGE Publications, Inc. <https://doi.org/10.4135/9781412985079>
- Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1), 40–49. <https://doi.org/10.1002/mpr.329>
- Bertsimas, D., Pawlowski, C., & Zhuo, Y. D. (2018). From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research*, 18(196), 1–39. <https://jmlr.org/papers/v18/17-073.html>
- Brilleman, S. L., Pachana, N. A., & Dobson, A. J. (2010). The impact of attrition on the representativeness of cohort studies of older people. *BMC Medical Research Methodology*, 10, 71–79. <https://doi.org/10.1186/1471-2288-10-71>
- Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24), 4279–4292. <https://doi.org/10.1002/sim.2673>
- Chatfield, M. D., Brayne, C. E., & Matthews, F. E. (2005). A systematic literature review of attrition between waves in longitudinal studies in the elderly shows a consistent pattern of dropout between differing studies. *Journal of Clinical Epidemiology*, 58(1), 13–19. <https://doi.org/10.1016/j.jclinepi.2004.05.006>
- Chen, L., Savalei, V., & Rhemtulla, M. (2023). Pay attention to the ignorable missing data mechanisms! An exploration of their impact on the efficiency of regression coefficients. *Multivariate Behavioral Research*, 58(6), 1134–1159. <https://doi.org/10.1080/00273171.2023.2193600>
- Choi, J. Y., Kyung, M., Hwang, H., & Park, J.-H. (2020). Bayesian Extended Redundancy Analysis: A Bayesian approach to component-based regression with dimension reduction. *Multivariate Behavioral Research*, 55(1), 30–48. <https://doi.org/10.1080/00273171.2019.1598837>

- Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330–351.
<https://doi.org/10.1037/1082-989X.6.4.330>
- Ditlhong, K. E., Ngesa, O. O., & Kombo, A. Y. (2018). A comparative analysis of generalized estimating equations methods for incomplete longitudinal ordinal data with ignorable dropouts. *Open Journal of Statistics*, 08(05), 770–792. <https://doi.org/10.4236/ojs.2018.85051>
- Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1), 54–75.
<https://doi.org/10.1214/ss/1177013815>
- Ekström, I., Josefsson, M., Bäckman, L., & Laukka, E. J. (2024). Predictors of cognitive aging profiles over 15 years: A longitudinal population-based study. *Psychology and Aging*, 39(5), 467–483.
<https://doi.org/10.1037/pag0000807>
- Fairclough, D. L. (2012). *Characterization of missing data* (2nd ed). Chapman and Hall/CRC.
<https://doi.org/10.1201/9781420061185>
- Feng, D., Cong, Z., & Silverstein, M. (2012). Missing data and attrition. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis* (1st ed., pp. 71–96). Routledge.
<https://doi.org/10.4324/9780203814208-4>
- Gabriel, K. R., & Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21(4), 489–498.
<https://doi.org/10.1080/00401706.1979.10489819>
- Gad, A. (2017). Imputation methods for longitudinal data: A comparative study. *International Journal of Statistical Distributions and Applications*, 3(4), 72–80.
<https://doi.org/10.11648/j.ijstd.20170304.13>
- Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60(1), 549–576. <https://doi.org/10.1146/annurev.psych.58.110405.085530>

- Guo, J., Wister, A., Wang, J., & Li, S. (2024). Life course risk and protective factors of multimorbidity resilience among older adults in rural China: A longitudinal study in Anhui province before and during COVID-19. *The Journals of Gerontology: Series B*, 79(4), 1–12.
<https://doi.org/10.1093/geronb/gbad196>
- Gupta, V. K., & Grover, G. (2017). Multiple imputation for gamma outcome variable using generalized linear model. *Journal of Statistical Computation and Simulation*, 87(10), 1980–1988.
<https://doi.org/10.1080/00949655.2017.1300904>
- He, L., Lu, Y., Falk, C. F., & Hwang, H. (2025). Evaluation of Generative Adversarial Imputation Nets' performance in handling missing data in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 32(6), 1026–1039.
<https://doi.org/10.1080/10705511.2025.2531528>
- Holmes Finch, W. (2018). Modeling high dimensional multilevel data using the lasso estimator: A simulation study. *Journal of Statistical and Econometric Methods*, 7(1), 51–75.
<https://uml.idm.oclc.org/login?url=https://www.proquest.com/scholarly-journals/modeling-high-dimensional-multilevel-data-using/docview/2573400935/se-2?accountid=14569>
- Huque, M. H., Carlin, J. B., Simpson, J. A., & Lee, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology*, 18(1), Article 168. <https://doi.org/10.1186/s12874-018-0615-6>
- Hwang, H., & Takane, Y. (2015). *Generalized Structured Component Analysis: A component-based approach to structural equation modeling*. CRC Press. <https://doi.org/10.1201/b17872>
- Ji, L., Chow, S.-M., Schermerhorn, A. C., Jacobson, N. C., & Cummings, E. M. (2018). Handling missing data in the modeling of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(5), 715–736. <https://doi.org/10.1080/10705511.2017.1417046>
- Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, 99(3), 422–431. <https://doi.org/10.1037/0033-2909.99.3.422>

- Kim, S., & Hwang, H. (2022). Evaluation of prediction-oriented model selection metrics for Extended Redundancy Analysis. *Frontiers in Psychology*, 13, Article 821897.
<https://doi.org/10.3389/fpsyg.2022.821897>
- Kim, S., Lee, S., Cardwell, R. L., Kim, Y., Park, T., & Hwang, H. (2020). An application of regularized Extended Redundancy Analysis via Generalized Estimating Equations to the study of co-occurring substance use among US adults. In D. Molenaar, J.-S. Kim, M. Wiberg, J. González, & U. Böckenholt (Eds.), *Quantitative psychology* (Vol. 322, pp. 365–376). Springer International Publishing AG. https://doi.org/10.1007/978-3-030-43469-4_27
- Kirkland, S. A., Griffith, L. E., Oz, U. E., Thompson, M., Wister, A., Kadowaki, L., Basta, N. E., McMillan, J., Wolfson, C., & Raina, P. (2023). Increased prevalence of loneliness and associated risk factors during the COVID-19 pandemic: Findings from the Canadian Longitudinal Study on Aging (CLSA). *BMC Public Health*, 23(1), 872–872. <https://doi.org/10.1186/s12889-023-15807-4>
- Lee, S., Kim, S., Kim, Y., Oh, B., Hwang, H., & Park, T. (2019). Pathway analysis of rare variants for the clustered phenotypes by using hierarchical structured components analysis. *BMC Medical Genomics*, 12(Suppl. 5), 100–109. <https://doi.org/10.1186/s12920-019-0517-4>
- Li, Y., Zhou, Q., Fan, Y., Pan, G., Dai, Z., & Lei, B. (2024). A novel machine learning-based imputation strategy for missing data in step-stress accelerated degradation test. *Heliyon*, 10(4), Article e26429. <https://doi.org/10.1016/j.heliyon.2024.e26429>
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22. <https://doi.org/10.1093/biomet/73.1.13>
- Mair, P & de Leeuw, J. (2009). Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software*, 31(4), 1–21. <https://doi.org/10.18637/jss.v031.i04>
- Meister, L. M., & Zahodne, L. B. (2022). Associations between social network components and cognitive domains in older adults. *Psychology and Aging*, 37(5), 591–603.
<https://doi.org/10.1037/pag0000672>

- O’Connell, M. E., Kadlec, H., Griffith, L. E., Maimon, G., Wolfson, C., Taler, V., Simard, M., Tuokko, H., Voll, S., Kirkland, S., & Raina, P. (2022). Methodological considerations when establishing reliable and valid normative data: Canadian Longitudinal Study on Aging (CLSA) neuropsychological battery. *The Clinical Neuropsychologist*, 36(8), 2168–2187.
<https://doi.org/10.1080/13854046.2021.1954243>
- Otse, E. J., Obunadike, G. N., & Abubakar, A. (2025). Linear regression approach to solving multicollinearity and overfitting in predictive analysis. *Journal of Science Research and Reviews*, 2(1), 108–117. <https://doi.org/10.70882/josrar.2025.v2i1.35>
- Park, H. L., O’Connell, J. E., & Thomson, R. G. (2003). A systematic review of cognitive decline in the general elderly population. *International Journal of Geriatric Psychiatry*, 18(12), 1121–1134.
<https://doi.org/10.1002/gps.1023>
- Quick, M., Rutter, E., Tyas, S. L., Maxwell, C., Law, J., & Oremus, M. (2022). Urban-rural variation in the association between social support availability and cognitive function in middle-aged and older adults: Results from the baseline Tracking Cohort of the Canadian Longitudinal Study on Aging. *Health & Place*, 77, Article 102894. <https://doi.org/10.1016/j.healthplace.2022.102894>
- Rahnenführer, J., De Bin, R., Benner, A., Ambrogi, F., Lusa, L., Boulesteix, A.-L., Migliavacca, E., Binder, H., Michiels, S., Sauerbrei, W., McShane, L., & for topic group “High-dimensional data” (TG9) of the STRATOS initiative. (2023). Statistical analysis of high-dimensional biomedical data: A gentle introduction to analytical goals, common approaches and challenges. *BMC Medicine*, 21(1), Article 182. <https://doi.org/10.1186/s12916-023-02858-y>
- Raina, P., Wolfson, C., Griffith, L., Kirkland, S., McMillan, J., Basta, N., Joshi, D., Oz, U. E., Sohel, N., Maimon, G., & Thompson, M. (2021). A longitudinal analysis of the impact of the COVID-19 pandemic on the mental health of middle-aged and older adults from the Canadian Longitudinal Study on Aging. *Nature Aging*, 1(12), 1137–1147. <https://doi.org/10.1038/s43587-021-00128-1>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
<https://doi.org/10.1093/biomet/63.3.581>

- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.
- Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC Medical Research Methodology*, 12(1), Article 46. <https://doi.org/10.1186/1471-2288-12-46>
- Statistics Canada. (2022). *A portrait of Canada's growing population aged 85 and older from the 2021 Census*. <https://www12.statcan.gc.ca/census-recensement/2021/as-sa/98-200-x/2021004/98-200-x2021004-eng.cfm>
- Steif, J., Brant, R., Sreepada, R. S., West, N., Murthy, S., & Görges, M. (2022). Prediction model performance with different imputation strategies: A simulation study using a north american ICU registry. *Pediatric Critical Care Medicine*, 23(1), e29–e44. <https://doi.org/10.1097/PCC.0000000000002835>
- Stinchcombe, A., Hammond, N. G., & Hopper, S. (2023). Changes in executive function in the Canadian longitudinal study on aging over 3-years: A focus on social determinants of health. *Frontiers in Psychology*, 14, Article 1060178. <https://doi.org/10.3389/fpsyg.2023.1060178>
- Takahashi, M. (2017). Statistical inference in missing data by MCMC and non-MCMC multiple imputation algorithms: Assessing the effects of between-imputation iterations. *Data Science Journal*, 16, Article 37. <https://doi.org/10.5334/dsj-2017-037>
- Takane, Y., & Hwang, H. (2005). An extended redundancy analysis and its applications to two practical examples. *Computational Statistics & Data Analysis*, 49(3), 785–808. <https://doi.org/10.1016/j.csda.2004.06.004>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2025). *Mice* [Computer software]. <https://CRAN.R-project.org/package=mice>
- van Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://doi.org/10.18637/jss.v045.i03>

- Vatcheva, K. P., Lee, M., McCormick, J. B., & Rahbar, M. H. (2016). Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale)*, 6(2), Article 227. <https://doi.org/10.4172/2161-1165.1000227>
- Wang, J. S., & Aronow, P. M. (2023). Listwise deletion in high dimensions. *Political Analysis*, 31(1), 149–155. <https://doi.org/10.1017/pan.2022.5>
- Wang, L., Zhou, J., & Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2), 353–360. <https://doi.org/10.1111/j.1541-0420.2011.01678.x>
- Wasito, I., & Mirkin, B. (2005). Nearest neighbour approach in the least-squares data imputation algorithms. *Information Sciences*, 169(1), 1–25. <https://doi.org/10.1016/j.ins.2004.02.014>
- Wasito, I., & Mirkin, B. (2006). Nearest neighbours in least-squares data imputation algorithms with different missing patterns. *Computational Statistics & Data Analysis*, 50(4), 926–949. <https://doi.org/10.1016/j.csda.2004.11.009>
- Woods, A. D., Gerasimova, D., Van Dusen, B., Nissen, J., Bainter, S., Uzdavines, A., Davis-Kean, P. E., Halvorson, M., King, K. M., Logan, J. A. R., Xu, M., Vasilev, M. R., Clay, J. M., Moreau, D., Joyal-Desmarais, K., Cruz, R. A., Brown, D. M. Y., Schmidt, K., & Elsherif, M. M. (2024). Best practices for addressing missing data through multiple imputation. *Infant and Child Development*, 33(1), Article e2407. <https://doi.org/10.1002/icd.2407>
- Yoo, W., Mayberry, R., Bae, S., Singh, K., Peter He, Q., & Lillard, J. (2014). A study of effects of multicollinearity in the multivariable analysis. *International Journal of Applied Science and Technology*, 4(5), 9–19. <https://ijast.thebrpi.org/journal/index/649>
- Yoon, J., Jordon, J., & van der Schaar, M. (2018). *GAIN: Missing data imputation using generative adversarial nets [Preprint]*. arXiv. <https://doi.org/10.48550/ARXIV.1806.02920>
- Zaninotto, P., Batty, G. D., Allerhand, M., & Deary, I. J. (2018). Cognitive function trajectories and their determinants in older people: 8 years of follow-up in the English Longitudinal Study of Ageing.

Journal of Epidemiology and Community Health, 72(8), 685–694. <https://doi.org/10.1136/jech-2017-210116>

Table 1*CLSA Empirical Application Variable Measures and Descriptive Statistics*

Variables	Description/measures	Mean (SD)	Missing (%)
Outcomes			
Cognition (T ₁)	Composite executive function score (baseline)	99.28 (15.42)	15.61
Cognition (T ₂)	Composite executive function score (follow-up 1)	101.55 (14.58)	30.26
Cognition (T ₃)	Composite executive function score (follow-up 2)	99.37 (15.36)	59.89
Predictors			
F1: SES			
Income	Less than \$20,000 (=1), \$50,000 (=2), \$100,000 (=3), \$150,000 (=4), or more (=5)	2.52 (1.02)	5.01
Education	No post-secondary degree (=1), trade certificate/diploma (=2), ..., bachelor's degree (=5), or above (=6)	4.06 (1.67)	15.10
Savings	Less than \$50,000 (=1), \$100,000 (=2), \$1,000,000 (=3), or more (=4)	2.50 (0.96)	15.51
F2: Health behaviours			
Diet	Number of daily fruit and vegetable servings: Less than two (=1), two (=2), ..., seven or more (=7)	4.02 (1.81)	5.28
Exercise	Weekly frequency of strength training: Never (=1), once or twice (=2), ..., five to seven times (=4)	1.57 (0.98)	4.41
Walking	Weekly frequency of outdoor walks: Never (=1), once or twice (=2), ..., five to seven times (=4)	3.07 (1.12)	4.45
F3: Social life			
Social outings	Annual frequency of social outings: Never (=1), once per year (=2), once per month, ..., once per day (=5)	3.50 (0.71)	0.19
Volunteering	Annual frequency of volunteering: Never (=1), once per year (=2), once per month, ..., once per day (=5)	2.31 (1.71)	0.20
Loneliness	Weekly frequency of feeling lonely: Never (=1), once or twice (=2), ..., five to seven times (=4)	1.39 (0.73)	0.40

Note. This table includes the measurement description, mean, standard deviation, and percent of cases with missing data in all variables included in the empirical analysis. CLSA = Canadian Longitudinal Study on Aging, SES = socioeconomic status, SD = standard deviation.

Table 2*Empirical Application Within-Predictor-Block Correlations*

	Income	Education	Savings	Diet	Exercise	Walking	Social outings	Volunteering	Loneliness
Income	1	0.29	0.40						
Education	0.29	1	0.23						
Savings	0.40	0.23	1						
Diet				1	0.12	0.13			
Exercise				0.12	1	0.07			
Walking				0.13	0.07	1			
Social outings							1	0.19	-0.10
Volunteering							0.19	1	-0.05
Loneliness							-0.10	-0.05	1

Note. This table demonstrates the Pearson correlations between predictors within the same block from the ERA-GEE model specified for the empirical application. ERA-GEE = Extended Redundancy Analysis-Generalized Estimating Equations.

Bolded values denote statistically significant Pearson correlations ($p < .05$).

Table 3*Empirical Application ERA-GEE Analysis Results*

Parameter	Variable	Missing-management method								
		LWD				MICE				LSI
		Estimate	95% CI [LL, UL]	SE	Estimate	95% CI [LL, UL]	SE	Estimate	95% CI [LL, UL]	SE
Weight coefficients	F1: SES									
	Income	0.158	[0.053, 0.258]	0.051	0.238	[0.183, 0.295]	0.029	0.333	[0.315, 0.352]	0.010
	Education	0.886	[0.828, 0.933]	0.027	0.798	[0.764, 0.831]	0.017	0.721	[0.706, 0.736]	0.008
	Savings	0.178	[0.074, 0.280]	0.053	0.238	[0.184, 0.291]	0.028	0.553	[0.535, 0.570]	0.009
	F2: Health behaviours									
	Diet	0.978	[0.872, 1.013]	0.037	0.952	[0.899, 0.987]	0.023	1.000	[0.985, 1.009]	0.006
	Exercise	-0.130	[-0.364, 0.119]	0.123	-0.089	[-0.214, 0.036]	0.064	0.154	[0.130, 0.169]	0.010
	Walking	0.134	[-0.131, 0.365]	0.126	0.228	[0.104, 0.350]	0.063	0.140	[0.074, 0.213]	0.039
	F3: Social life									
	Social outings	0.357	[-0.211, 0.785]	0.255	0.262	[0.286, 0.564]	0.071	0.440	[0.375, 0.506]	0.035
Regression coefficients	Volunteering	0.793	[0.250, 0.985]	0.187	0.288	[0.348, 0.615]	0.068	0.456	[0.380, 0.522]	0.037
	Loneliness	-0.330	[-0.780, 0.231]	0.254	-0.385	[-0.747, -0.516]	0.059	-0.698	[-0.757, -0.627]	0.033
	Y1: Cognition (T1)									
	SES	0.241	[0.216, 0.267]	0.013	0.235	[0.223, 0.247]	0.006	0.549	[0.539, 0.559]	0.005
	Health behaviours	0.084	[0.059, 0.110]	0.013	0.085	[0.073, 0.097]	0.006	0.294	[0.276, 0.314]	0.010
	Social life	0.051	[0.026, 0.078]	0.013	0.045	[0.063, 0.088]	0.006	0.098	[0.089, 0.108]	0.005

Parameter	Variable	Missing-management method								
		LWD			MICE			LSI		
		Estimate	95% CI [LL, UL]	SE	Estimate	95% CI [LL, UL]	SE	Estimate	95% CI [LL, UL]	SE
	Y2: Cognition (T2)									
	SES	0.232	[0.208, 0.256]	0.012	0.226	[0.214, 0.239]	0.006	0.61	[0.598, 0.623]	0.006
	Health behaviours	0.093	[0.068, 0.120]	0.013	0.090	[0.078, 0.102]	0.006	0.249	[0.234, 0.266]	0.008
	Social life	0.033	[0.014, 0.062]	0.012	0.041	[0.057, 0.082]	0.006	0.085	[0.076, 0.093]	0.004
	Y3: Cognition (T3)									
	SES	0.233	[0.209, 0.257]	0.013	0.237	[0.223, 0.250]	0.007	0.706	[0.694, 0.718]	0.006
	Health behaviours	0.094	[0.070, 0.120]	0.013	0.086	[0.074, 0.099]	0.007	0.267	[0.252, 0.284]	0.008
	Social life	0.0416	[0.021, 0.069]	0.012	0.042	[0.058, 0.083]	0.006	0.112	[0.101, 0.123]	0.005

Note. This table demonstrates the ERA-GEE parameter estimates with their corresponding confidence intervals and standard errors under each missing-management method. ERA-GEE = Extended Redundancy Analysis-Generalized Estimating Equations, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation, CI = confidence interval, LL = lower limit, UL = upper limit, SE = standard error, SES = socioeconomic status, T1 = baseline, T2 = follow-up 1, T3 = follow-up 3. Bolded values represent statistically significant parameter values where the corresponding 95% confidence interval does not include zero.

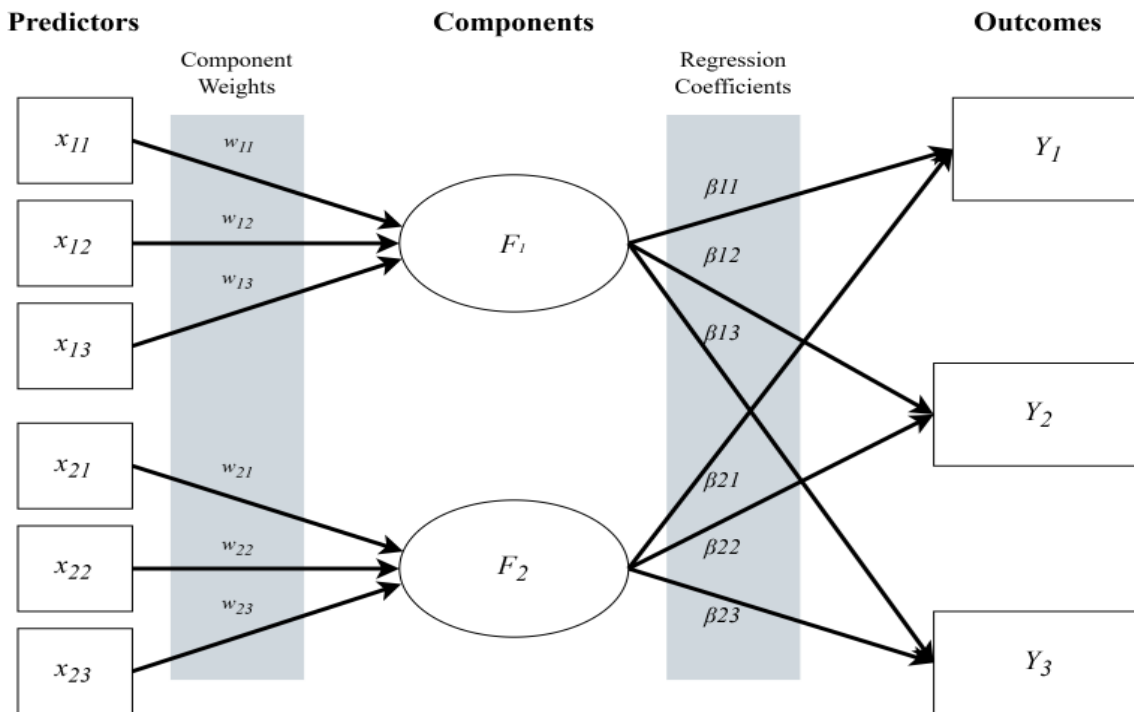
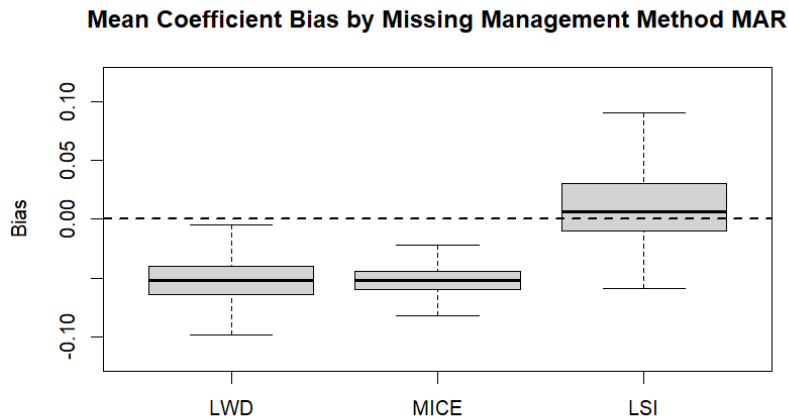
Figure 1*Schematic Diagram of ERA-GEE Model Used for Simulated Data Generation*

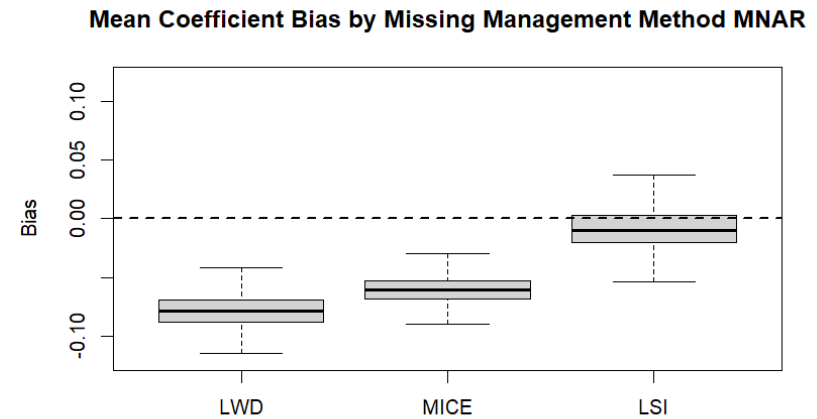
Figure 2

Regression Coefficient Bias Across Missing-Management Methods under MAR and MNAR

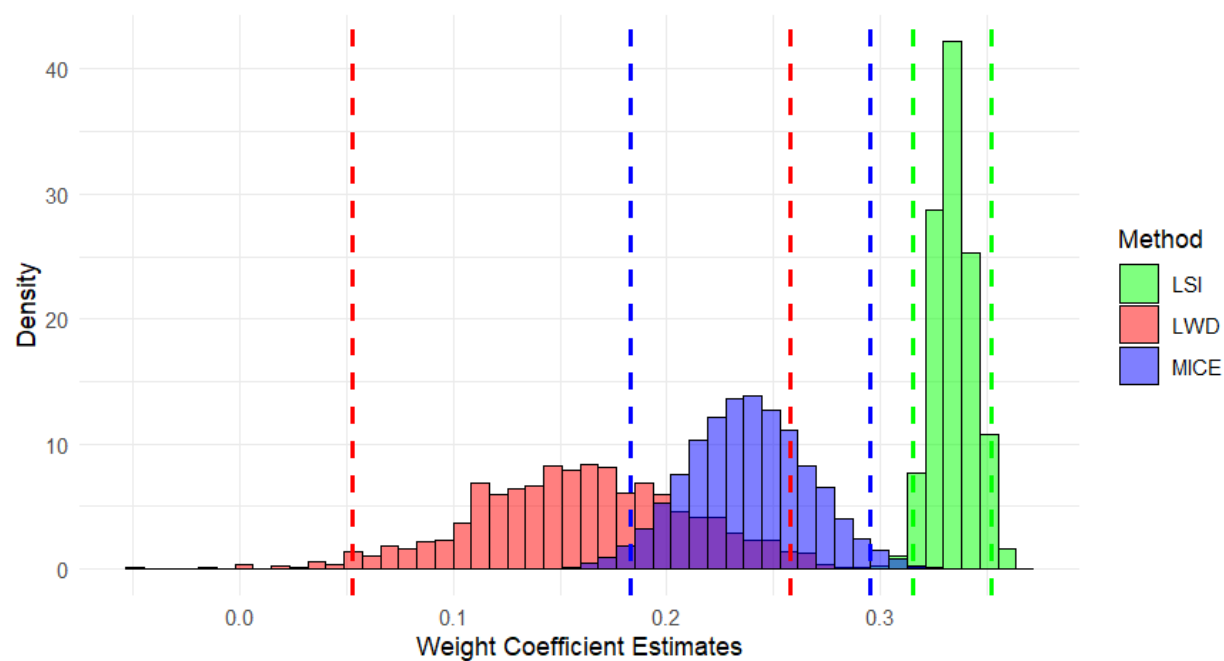
A)



B)



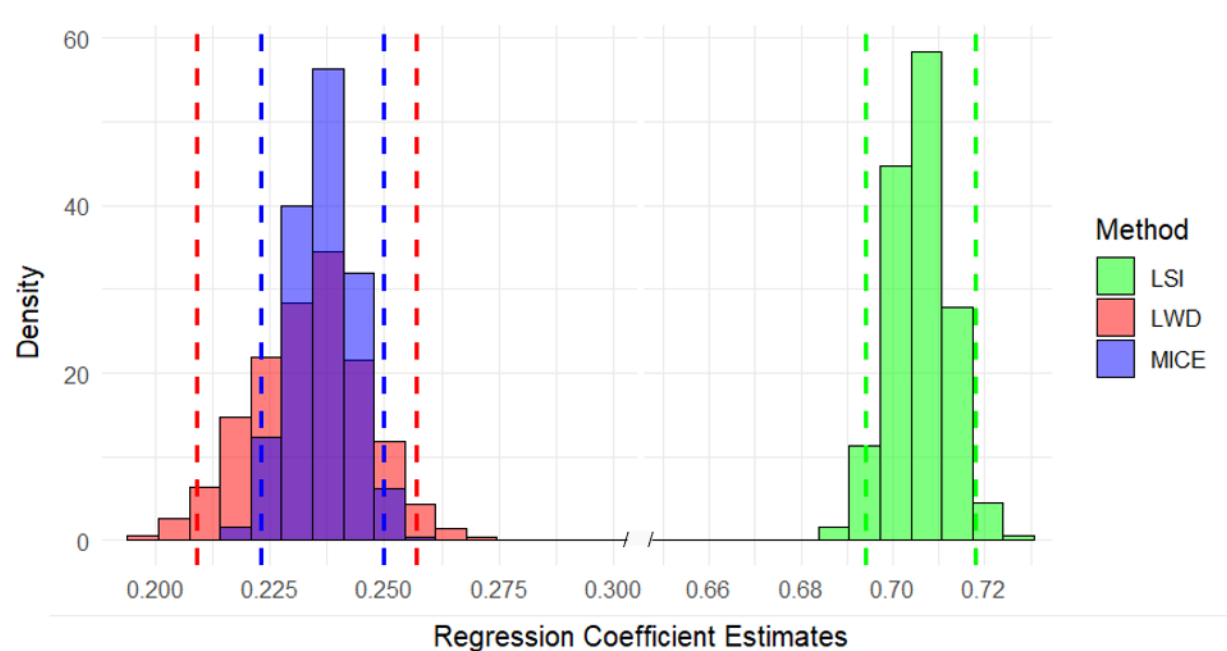
Note. This figure displays boxplots of bias values across 1,000 iterations in the missing-data conditions where 10% of values are missing, the sample size is 500, and the missing mechanism is MAR (left side, A) or MNAR (right side, B). In the boxplots, solid black horizontal lines represent the median, the grey boxes represent the interquartile range, and the whiskers represent the minimum and maximum values excluding outliers. The dashed horizontal lines represent a bias value of zero indicating perfect accuracy. LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation, MAR = missing at random, MNAR = missing not at random.

Figure 3*Bootstrapped Component Weight Distributions Across Missing-Management Methods*

Note. This figure shows three density plots from 1,000 estimates of the component weight for income for each missing-management method on the same x-axis. The vertical dashed lines demonstrate the 95% confidence intervals for each method. LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Figure 4

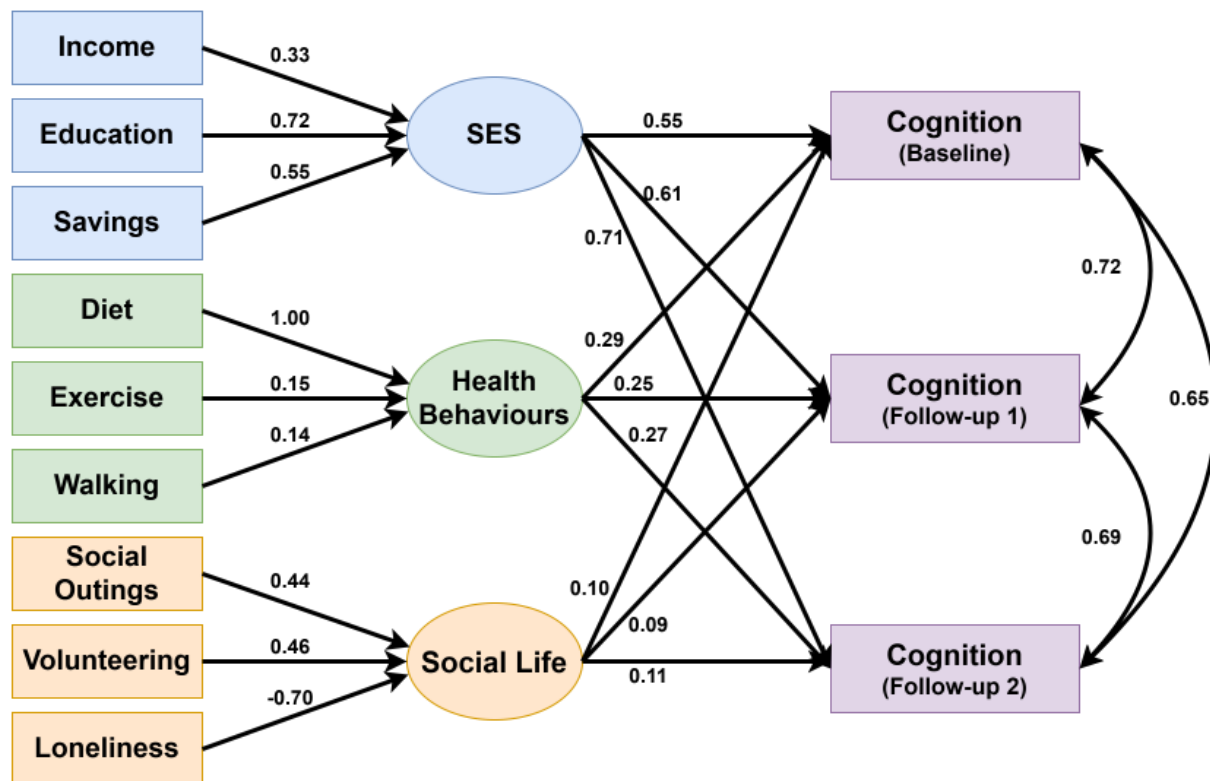
Bootstrapped Regression Coefficient Distributions Across Missing-Management Methods



Note. This figure shows three density plots from 1,000 estimates of the regression coefficient for the impact of socioeconomic status on cognition at follow-up 2 for each missing-management method on the same x-axis. The vertical dashed lines demonstrate the 95% confidence intervals for each method. The x-axis is split at 0.3 and 0.65 to minimize the amount of empty space in the figure. LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Figure 5

Schematic of ERA-GEE Model from the Empirical Application with LSI-ERA-GEE Results



Note. Values on the left represent component weights, values in the middle represent regression coefficients, and values on the right represent Pearson correlations. ERA-GEE = Extended Redundancy Analysis-Generalized Estimating Equations, LSI-ERA-GEE = Least-squares Imputation-Extended Redundancy Analysis-Generalized Estimating Equations, SES = socioeconomic status.

Bolded weights and coefficients represent statistically significant values where the corresponding 95% confidence intervals does not include zero. Bolded Pearson correlation values are statistically significant ($p < .05$).

Appendix A

LSI Missing Value Estimating Equation Derivations

This appendix provides the algebraic details for how the estimating equations in the main text for imputing missing predictors and outcomes are derived from the ERA-GEE least-squares objective function. Most of the notation in this appendix is identical to the main text; any new notation is defined as needed. The derivations presented here assume a quadratic loss formulation under an identity-link specification, consistent with the Gaussian working approximation adopted in the estimation procedure. Accordingly, these derivations do not directly extend to general exponential family settings with non-identity link functions. For notational simplicity, the subscript i is omitted from Σ , as the covariance matrix is assumed to be common across individuals.

Estimating Equation for Missing Outcome Values

For predicting missing outcomes via LSI, the process begins with ERA-GEE's standard objective function:

$$\Phi = \sum_{i=1} (y_i - \mu_i)' \Sigma^{-1} (y_i - \mu_i),$$

where μ_i is the linear predictor of the outcome vector calculated as $\mu_i = x_i \mathbf{W} \mathbf{B}$ and y_i is the observed outcome vector which is equivalent to $\tilde{z}_i$ from the main text as all outcomes are Gaussian in the present setting. During this step, x_i , $\mathbf{W}$, $\mathbf{B}$, and Σ are treated as fixed. Next, all elements connected to the outcome vector are decomposed into missing and non-missing elements as follows:

$$y_i = \begin{pmatrix} y_{i-j} \\ y_{ij} \end{pmatrix}, \mu_i = \begin{pmatrix} \mu_{i-j} \\ \mu_{ij} \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{-j-j} & \Sigma_{-jj} \\ \Sigma_{j-j} & \Sigma_{jj} \end{pmatrix}, \Omega = \Sigma^{-1} = \begin{pmatrix} \Omega_{-j-j} & \Omega_{-jj} \\ \Omega_{j-j} & \Omega_{jj} \end{pmatrix},$$

where the j subscript indicates the column that currently contains missing values, the $-j$ subscript indicates the non-missing columns, and $\mathbf{\Omega}$ denotes the inverse covariance matrix. The covariance matrix is decomposed into covariances between observed outcomes, between observed and missing outcomes, and within the missing outcome. In the next step, the original objective function is expanded with the decomposition of missing and non-missing columns and restricted to individuals with missing values on the j th outcome. The expanded objective function becomes:

$$\Phi = \sum_{i \in Y_{ind,j=0}} (y_{i-j} - \mu_{i-j})' \mathbf{\Omega}_{-j-j} (y_{i-j} - \mu_{i-j}) + 2(y_{ij} - \mu_{ij})' \mathbf{\Omega}_{-jj}' (y_{i-j} - \mu_{i-j}) + (y_{ij} - \mu_{ij})' \mathbf{\Omega}_{jj} (y_{ij} - \mu_{ij}).$$

This is the least-squares subproblem that is solved to estimate missing values in the j th column of y . At this step, everything except y_{ij} is fixed. Taking the derivative with respect to y_{ij} and setting it equal to zero:

$$\frac{\Phi}{\Phi y_{ij}} = 2\mathbf{\Omega}_{jj}(y_{ij} - \mu_{ij}) + 2\mathbf{\Omega}_{j-j}'(y_{i-j} - \mu_{i-j}) = 0.$$

Solving for y_{ij} then yields:

$$\hat{y}_{ij} = \mu_{ij} + \mathbf{\Sigma}_{j-j} \mathbf{\Omega}_{-j-j} (y_{i-j} - \mu_{i-j}).$$

Solving this equation produces the value of $\hat{y}_{ij}$ that minimizes ERA-GEE's least-squares objective function. Although in the final equation missing values are estimated at the row level, the subproblem can be solved in matrix form using the outcome matrices $\hat{\mathbf{Y}}_j$ and μ_j . Since this procedure is carried out in a column-wise manner and the covariance matrix is assumed to be common across all observations, the missing outcomes in each column can be stacked across observations with missing data, allowing all missing values in each column to be imputed at once.

This update can be interpreted as a conditional expectation under a multivariate normal working model. Specifically, the imputed value adjusts the model-based prediction μ_{ij} using the residual structure of the observed outcomes, weighted by the covariance matrix. Thus, missing outcomes are imputed in a way that is consistent with both the current model fit and the observed dependence structure among outcomes.

Estimating Equation for Missing Predictor Values

For estimating missing predictors via LSI in ERA-GEE, the same general procedure is followed. The process for predicting missing predictor values also begins with the original least squares objective function:

$$\Phi = \sum_{i=1} (y_i - x_i \mathbf{M})' \boldsymbol{\Sigma}^{-1} (y_i - x_i \mathbf{M}),$$

where $\mathbf{M}$ is a P by Q matrix that connects each predictor to each outcome and is defined as $\mathbf{M} = \mathbf{W}\mathbf{B}$. When estimating missing predictor values, y_i , $\mathbf{M}$, and $\boldsymbol{\Sigma}$ are treated as fixed. Next, the linear predictor $x_i \mathbf{M}$ is decomposed into missing and non-missing elements as follows:

$$x_i \mathbf{M} = x_{ij} \mathbf{M}_j + x_{i-j} \mathbf{M}_{-j},$$

where the j subscript indicates the column of x that currently contains missing values or the row of $\mathbf{M}$ pertaining to the j th predictor. The $-j$ subscript indicates all other columns and rows of x and $\mathbf{M}$, respectively, that contain non-missing values. Restricting to individuals with missing values on the j th predictor gives:

$$\Phi = \sum_{i \in X_{ind,j=0}} (y_i - x_{i-j} \mathbf{M}_{-j} - x_{ij} \mathbf{M}_j)' \boldsymbol{\Sigma}^{-1} (y_i - x_{i-j} \mathbf{M}_{-j} - x_{ij} \mathbf{M}_j).$$

This is the least-squares subproblem that is solved to estimate missing values in the j th column of x . At this step, everything except x_{ij} is fixed. Taking the derivative with respect to x_{ij} and setting it equal to zero:

$$\frac{\Phi}{\Phi x_{ij}} = -2\mathbf{M}_j\boldsymbol{\Sigma}^{-1}(y_i - x_{i-j}\mathbf{M}_{-j})' + 2x_{ij}\mathbf{M}_j\boldsymbol{\Sigma}^{-1}\mathbf{M}_j' = 0.$$

Solving for x_{ij} then yields:

$$\hat{x}_{ij} = \frac{\mathbf{M}_j\boldsymbol{\Sigma}^{-1}(y_i - x_{i-j}\mathbf{M}_{-j})'}{\mathbf{M}_j\boldsymbol{\Sigma}^{-1}\mathbf{M}_j'}.$$

Solving this equation produces the value of $\hat{x}_{ij}$ that minimizes ERA-GEE's least-squares objective function. Similar to the estimating equation for outcomes, missing predictor values in each column can be stacked across observations with missing data, allowing all missing values in each column to be imputed at once using the predictor and outcome matrices $\hat{\mathbf{X}}_j$ and $\mathbf{Y}$, respectively.

This update corresponds to the coefficient obtained by projecting the partial residual vector $(y_i - x_{i-j}\mathbf{M}_{-j})$ onto the direction defined by $\mathbf{M}_j$, under the inner product induced by $\boldsymbol{\Sigma}^{-1}$. Importantly, this interpretation highlights that missing predictor values are not imputed based on relationships between predictors alone. Instead, they are determined so as to optimally explain the current outcome residual under the fitted model. In this sense, the procedure performs model-consistent imputation, aligning the imputed predictor values with the current regression structure defined by $\mathbf{M}=\mathbf{WB}$.

Appendix B

Simulation Feasibility Results Tables

Table B1

Convergence Probabilities Across Missing-Data Conditions Under LWD

Mechanism	<i>N</i>	Missing proportion			
		5%	10%	20%	30%
MCAR	50	1.00	1.00	0.87	0.03
	100	1.00	1.00	1.00	0.33
	300	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00
MAR	50	1.00	1.00	0.97	0.24
	100	1.00	1.00	1.00	0.81
	300	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00
MNAR	50	1.00	1.00	1.00	1.00
	100	1.00	1.00	1.00	1.00
	300	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00

Note. LWD = listwise deletion, MCAR = missing completely at random, MAR = missing at random, MNAR = missing not at random.

Table B2*Proportion of Retained Sample Across Missing-Data Conditions Following LWD*

Mechanism	N	Missing proportion			
		5%	10%	20%	30%
MCAR	50	0.69	0.39	0.14	0.11
	100	0.63	0.39	0.13	0.06
	300	0.63	0.39	0.13	0.04
	500	0.63	0.39	0.13	0.04
MAR	50	0.70	0.41	0.17	0.11
	100	0.64	0.41	0.16	0.07
	300	0.64	0.40	0.16	0.06
	500	0.64	0.40	0.16	0.06
MNAR	50	0.79	0.59	0.36	0.22
	100	0.75	0.59	0.36	0.22
	300	0.75	0.59	0.36	0.22
	500	0.75	0.59	0.36	0.22

Note. This table demonstrates the proportion of the original sample that remained after all cases with any missing data were removed prior to analysis when LWD was the missing-management method. LWD = listwise deletion, MCAR = missing completely at random, MAR = missing at random, MNAR = missing not at random.

Table B3*Mean Convergence Time in All Simulation Conditions*

Mechanism	N	Missing proportion											
		5%			10%			20%			30%		
		LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI
MCAR	50	0.03	2.44	0.29	0.02	2.47	1.10	0.09	2.56	2.49	0.20	2.82	3.66
	100	0.04	3.17	0.64	0.03	3.23	2.06	0.02	3.33	4.71	0.17	3.48	6.15
	300	0.11	5.86	2.23	0.07	5.97	6.65	0.03	6.16	15.62	0.03	6.23	17.22
	500	0.16	8.47	3.83	0.11	8.63	12.76	0.05	8.93	28.92	0.02	9.04	28.97
MAR	50	0.03	2.44	0.33	0.02	2.46	1.12	0.06	2.55	2.74	0.18	2.82	3.74
	100	0.04	3.17	0.68	0.03	3.24	2.11	0.02	3.32	4.73	0.09	3.48	6.43
	300	0.11	5.86	2.55	0.07	6.05	7.64	0.04	6.15	16.19	0.02	6.23	17.17
	500	0.16	8.50	4.01	0.11	8.68	13.94	0.05	8.92	28.57	0.02	9.05	29.18
MNAR	50	0.03	2.44	0.38	0.03	2.49	0.84	0.03	2.59	1.70	0.03	2.97	2.82
	100	0.05	3.23	0.95	0.04	3.27	1.57	0.03	3.36	3.04	0.02	2.77	4.14
	300	0.12	5.95	2.58	0.10	6.10	5.58	0.07	6.24	11.21	0.03	5.46	13.19
	500	0.18	8.59	4.74	0.15	8.78	9.81	0.10	9.08	21.30	0.05	8.31	25.40

Note. This table represents the mean convergence time in seconds for each condition. LWD = listwise deletion, MCAR = missing completely at random, MAR = missing at random, MNAR = missing not at random, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Table B4*Mean Number of Iterations to Convergence in All Simulation Conditions*

Mechanism	N	Missing proportion											
		5%			10%			20%			30%		
		LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI
MCAR	50	6.96	6.61	25.10	8.27	6.88	91.13	84.33	7.50	225.25	190.00	8.70	313.38
	100	6.25	6.04	32.24	6.82	6.12	102.49	12.50	6.39	246.22	172.64	6.83	303.63
	300	5.64	5.48	42.22	5.86	5.56	126.89	6.75	5.69	297.29	19.66	5.84	314.98
	500	5.42	5.23	44.63	5.63	5.30	149.63	6.23	5.42	317.68	8.41	5.57	316.37
MAR	50	6.96	6.62	27.84	8.16	6.86	94.35	51.62	7.47	238.49	185.16	8.74	322.42
	100	6.22	6.03	34.34	6.81	6.13	104.94	9.75	6.38	237.28	98.51	6.81	318.97
	300	5.64	5.49	48.34	5.86	5.55	138.34	6.64	5.68	296.78	9.19	5.83	314.91
	500	5.39	5.25	46.88	5.64	5.28	154.66	6.12	5.42	315.16	7.48	5.57	321.23
MNAR	50	6.95	6.64	32.40	7.71	6.97	70.79	10.31	7.73	145.58	21.20	9.53	239.84
	100	6.25	6.06	43.21	6.56	6.18	74.75	7.50	6.52	144.97	9.65	7.13	251.66
	300	5.68	5.52	45.57	5.82	5.62	103.05	6.10	5.79	206.17	6.70	5.96	278.03
	500	5.43	5.28	52.59	5.61	5.35	110.96	5.87	5.52	249.11	6.20	5.69	314.99

Note. MCAR = missing completely at random, MAR = missing at random, MNAR = missing not at random, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Table B5*Mean Total-Loss Value at Convergence in All Simulation Conditions*

Mechanism	N	Missing proportion											
		5%			10%			20%			30%		
		LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI
MCAR	50	0.6159	0.6161	0.5435	0.5888	0.6102	0.4327	0.3254	0.6205	0.2671	0.1198	0.6613	0.1430
	100	0.6312	0.6310	0.5431	0.6200	0.6313	0.4494	0.5317	0.6302	0.2907	0.2142	0.6376	0.1668
	300	0.6368	0.6427	0.5505	0.6359	0.6387	0.4622	0.6203	0.6411	0.3046	0.5206	0.6415	0.1836
	500	0.6398	0.6409	0.5510	0.6343	0.6425	0.4660	0.6296	0.6408	0.3076	0.5827	0.6478	0.1884
MAR	50	0.6338	0.6138	0.5444	0.6288	0.6169	0.4387	0.4502	0.6232	0.2793	0.1959	0.6872	0.1629
	100	0.6462	0.6294	0.5390	0.6585	0.6316	0.4523	0.6420	0.6317	0.2968	0.3838	0.6443	0.1863
	300	0.6505	0.6376	0.5498	0.6667	0.6384	0.4616	0.6897	0.6405	0.3110	0.6764	0.6541	0.2006
	500	0.6546	0.6398	0.5529	0.6695	0.6424	0.4660	0.6934	0.6428	0.3130	0.7087	0.6510	0.2015
MNAR	50	0.7897	0.6796	0.6003	0.8918	0.7861	0.5748	0.9801	0.9600	0.5330	0.9017	1.1490	0.4734
	100	0.8343	0.7184	0.6088	0.9264	0.8020	0.5957	1.0471	0.9769	0.5676	1.0899	1.1587	0.5141
	300	0.8487	0.7290	0.6222	0.9439	0.8171	0.6118	1.0730	0.9936	0.5834	1.1779	1.1821	0.5488
	500	0.8531	0.7330	0.6263	0.9462	0.8185	0.6129	1.0871	0.9948	0.5887	1.1839	1.1874	0.5510

Note. MCAR = missing completely at random, MAR = missing at random, MNAR = missing not at random, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Appendix C

Simulation Estimator Behaviour Results Tables

Table C1

Variance Results: MCAR Simulation Conditions

Parameter	N	Missing proportion											
		5%			10%			20%			30%		
		LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI
Coefficients	50	0.0067	0.0039	0.0052	0.0147	0.0047	0.0119	189.4691	0.0058	0.0448	3167.327	0.0080	0.1031
	100	0.0037	0.0019	0.0029	0.0070	0.0022	0.0060	0.1589	0.0026	0.0222	357.8578	0.0035	0.0475
	300	0.0011	0.0006	0.0009	0.0021	0.0007	0.0023	0.0068	0.0008	0.0085	1.5246	0.0010	0.0153
	500	0.0007	0.0004	0.0005	0.0012	0.0004	0.0016	0.0041	0.0005	0.0053	0.0288	0.0006	0.0098
Weights	50	0.0125	0.0097	0.0103	0.0293	0.0129	0.0157	0.9609	0.0219	0.0296	60.5013	0.0386	0.0473
	100	0.0062	0.0046	0.0052	0.0103	0.0055	0.0081	0.0896	0.0081	0.0124	2.1588	0.0140	0.0165
	300	0.0019	0.0014	0.0018	0.0031	0.0016	0.0033	0.0103	0.0023	0.0048	0.1416	0.0036	0.0052
	500	0.0011	0.0008	0.0010	0.0018	0.0010	0.0022	0.0061	0.0014	0.0030	0.0299	0.0021	0.0031

Note. This table displays the mean variance for pooled regression coefficients and pooled component weights across 1,000 iterations in all simulation conditions where the missing mechanism is **MCAR**. MCAR = missing completely at random, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Table C2*Variance Results: MAR Simulation Conditions*

Parameter	N	Missing proportion											
		5%			10%			20%			30%		
		LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI
Coefficients	50	0.0067	0.0041	0.0053	0.0150	0.0046	0.0133	70.8744	0.0061	0.0495	246.0564	0.0084	0.0949
	100	0.0036	0.0020	0.0027	0.0063	0.0022	0.0064	0.0236	0.0027	0.0206	590.2069	0.0036	0.0442
	300	0.0011	0.0006	0.0009	0.0020	0.0007	0.0025	0.0058	0.0008	0.0083	0.0230	0.0010	0.0145
	500	0.0007	0.0004	0.0006	0.0012	0.0004	0.0016	0.0036	0.0005	0.0052	0.0113	0.0006	0.0092
Weights	50	0.0129	0.0094	0.0106	0.0282	0.0120	0.0161	0.5730	0.0222	0.0283	11.1083	0.0366	0.0425
	100	0.0064	0.0044	0.0053	0.0108	0.0055	0.0082	0.0507	0.0082	0.0120	0.8391	0.0133	0.0165
	300	0.0019	0.0014	0.0019	0.0032	0.0017	0.0035	0.0090	0.0024	0.0046	0.0438	0.0035	0.0049
	500	0.0011	0.0008	0.0011	0.0019	0.0010	0.0023	0.0051	0.0014	0.0032	0.0179	0.0021	0.0029

Note. This table displays the mean variance for pooled regression coefficients and pooled component weights across 1,000 iterations in all simulation conditions where the missing mechanism is **MAR**. MAR = missing at random, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Table C3*Variance Results: MNAR Simulation Conditions*

Parameters	N	Missing proportion											
		5%			10%			20%			30%		
		LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI
Coefficients	50	0.0070	0.0044	0.0048	0.0111	0.0057	0.0082	0.0269	0.0087	0.0202	9.6363	0.0140	0.0542
	100	0.0035	0.0022	0.0025	0.0052	0.0026	0.0041	0.0101	0.0038	0.0094	0.0221	0.0059	0.0229
	300	0.0011	0.0007	0.0008	0.0017	0.0008	0.0014	0.0034	0.0013	0.0039	0.0060	0.0019	0.0086
	500	0.0007	0.0004	0.0005	0.0009	0.0005	0.0009	0.0018	0.0007	0.0028	0.0036	0.0010	0.0054
Weights	50	0.0123	0.0110	0.0106	0.0201	0.0152	0.0173	0.0512	0.0256	0.0390	0.1915	0.0499	0.0714
	100	0.0060	0.0052	0.0048	0.0086	0.0067	0.0074	0.0183	0.0106	0.0122	0.0396	0.0177	0.0244
	300	0.0019	0.0016	0.0016	0.0026	0.0021	0.0023	0.0052	0.0032	0.0037	0.0096	0.0053	0.0053
	500	0.0011	0.0009	0.0010	0.0015	0.0012	0.0015	0.0028	0.0019	0.0026	0.0055	0.0030	0.0034

Note. This table displays the mean variance for pooled regression coefficients and pooled component weights across 1,000 iterations in all simulation conditions where the missing mechanism is **MNAR**. MNAR = missing not at random, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Table C4*Bias Results: MCAR Simulation Conditions*

Parameter	N	Missing proportion											
		5%			10%			20%			30%		
		LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI
Coefficients	50	-0.0450	-0.0471	-0.0303	-0.0389	-0.0467	0.0181	2.2717	-0.0441	0.1584	5.1931	-0.0440	0.2495
	100	-0.0488	-0.0489	-0.0280	-0.0452	-0.0512	0.0141	-0.0132	-0.0492	0.1456	4.7430	-0.0486	0.2547
	300	-0.0522	-0.0512	-0.0325	-0.0507	-0.0513	0.0114	-0.0480	-0.0516	0.1497	0.0547	-0.0519	0.2504
	500	-0.0522	-0.0523	-0.0328	-0.0520	-0.0525	0.0121	-0.0506	-0.0516	0.1538	-0.0386	-0.0524	0.2455
Weights	50	-0.1782	-0.1784	-0.1674	-0.1810	-0.1789	-0.1434	-0.2522	-0.1792	-0.1012	-0.3609	-0.1827	-0.0795
	100	-0.1782	-0.1782	-0.1650	-0.1788	-0.1780	-0.1431	-0.1905	-0.1780	-0.1007	-0.3069	-0.1791	-0.0743
	300	-0.1782	-0.1782	-0.1651	-0.1784	-0.1782	-0.1421	-0.1786	-0.1781	-0.0982	-0.1966	-0.1780	-0.0801
	500	-0.1781	-0.1779	-0.1647	-0.1784	-0.1780	-0.1399	-0.1781	-0.1780	-0.0965	-0.1823	-0.1779	-0.0807

Note. This table displays the mean bias for pooled regression coefficients and pooled component weights across 1,000 iterations in all simulation conditions where the missing mechanism is **MCAR**. MCAR = missing completely at random, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Table C5*Bias Results: MAR Simulation Conditions*

Parameter	N	Missing proportion											
		5%			10%			20%			30%		
		LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI
Coefficients	50	-0.0467	-0.0481	-0.0283	-0.0380	-0.0463	0.0233	0.9333	-0.0418	0.1650	4.3962	-0.0402	0.2612
	100	-0.0498	-0.0497	-0.0265	-0.0468	-0.0502	0.0165	-0.0409	-0.0481	0.1378	3.3283	-0.0458	0.2460
	300	-0.0519	-0.0520	-0.0312	-0.0512	-0.0521	0.0117	-0.0502	-0.0509	0.1469	-0.0401	-0.0510	0.2368
	500	-0.0524	-0.0529	-0.0326	-0.0513	-0.0516	0.0139	-0.0503	-0.0525	0.1478	-0.0467	-0.0524	0.2402
Weights	50	-0.1775	-0.1787	-0.1671	-0.1778	-0.1791	-0.1412	-0.2289	-0.1785	-0.1021	-0.3066	-0.1799	-0.0768
	100	-0.1768	-0.1781	-0.1643	-0.1764	-0.1780	-0.1415	-0.1786	-0.1779	-0.1026	-0.2584	-0.1775	-0.0775
	300	-0.1771	-0.1780	-0.1642	-0.1763	-0.1781	-0.1403	-0.1744	-0.1779	-0.0991	-0.1772	-0.1776	-0.0832
	500	-0.1772	-0.1778	-0.1644	-0.1760	-0.1779	-0.1393	-0.1746	-0.1776	-0.0978	-0.1732	-0.1775	-0.0830

Note. This table displays the mean bias for pooled regression coefficients and pooled component weights across 1,000 iterations in all simulation conditions where the missing mechanism is **MAR**. MAR = missing at random, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Table C6*Bias Results: MNAR Simulation Conditions*

Parameter	N	Missing proportion											
		5%			10%			20%			30%		
		LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI
Coefficients	50	-0.0556	-0.0512	-0.0307	-0.0702	-0.0537	0.0023	-0.0834	-0.0561	0.0850	0.1271	-0.0693	0.1851
	100	-0.0621	-0.0557	-0.0298	-0.0743	-0.0577	-0.0012	-0.0910	-0.0614	0.0712	-0.1004	-0.0779	0.1812
	300	-0.0667	-0.0566	-0.0329	-0.0784	-0.0599	-0.0054	-0.0982	-0.0697	0.0701	-0.1129	-0.0839	0.1656
	500	-0.0667	-0.0571	-0.0337	-0.0785	-0.0602	-0.0063	-0.0978	-0.0696	0.0748	-0.1149	-0.0850	0.1688
Weights	50	-0.1658	-0.1750	-0.1666	-0.1597	-0.1700	-0.1499	-0.1537	-0.1626	-0.1250	-0.1862	-0.1653	-0.1078
	100	-0.1641	-0.1731	-0.1623	-0.1570	-0.1697	-0.1494	-0.1493	-0.1609	-0.1232	-0.1471	-0.1545	-0.0972
	300	-0.1630	-0.1733	-0.1623	-0.1563	-0.1686	-0.1490	-0.1464	-0.1596	-0.1244	-0.1396	-0.1508	-0.1034
	500	-0.1634	-0.1731	-0.1620	-0.1564	-0.1685	-0.1486	-0.1456	-0.1599	-0.1226	-0.1384	-0.1508	-0.1028

Note. This table displays the mean bias for pooled regression coefficients and pooled component weights across 1,000 iterations in all simulation conditions where the missing mechanism is **MNAR**. MNAR = missing not at random, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Table C7*RMSE Results: MCAR Simulation Conditions*

Parameter	N	Missing proportion											
		5%			10%			20%			30%		
		LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI
Coefficients	50	0.0991	0.0869	0.0854	0.1279	0.0908	0.1050	2.7784	0.0947	0.2248	13.0037	0.1055	0.3431
	100	0.0862	0.0769	0.0725	0.1004	0.0795	0.0826	0.1857	0.0815	0.1873	6.1950	0.0853	0.3043
	300	0.0741	0.0705	0.0619	0.0788	0.0713	0.0646	0.1008	0.0722	0.1743	0.2814	0.0738	0.2761
	500	0.0717	0.0701	0.0595	0.0748	0.0703	0.0607	0.0891	0.0701	0.1742	0.1374	0.0714	0.2678
Weights	50	0.2123	0.2064	0.1982	0.2462	0.2140	0.1930	0.7378	0.2316	0.1968	2.6063	0.2635	0.2223
	100	0.1971	0.1936	0.1840	0.2077	0.1958	0.1748	0.3188	0.2016	0.1575	1.0846	0.2160	0.1569
	300	0.1869	0.1857	0.1755	0.1903	0.1858	0.1607	0.2075	0.1879	0.1331	0.3673	0.1913	0.1211
	500	0.1848	0.1838	0.1730	0.1869	0.1843	0.1555	0.1976	0.1853	0.1262	0.2466	0.1872	0.1144

Note. This table displays the mean RMSE for pooled regression coefficients and pooled component weights across 1,000 iterations in all simulation conditions where the missing mechanism is **MCAR**. RMSE = root mean squared error, MCAR = missing completely at random, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Table C8*RMSE Results: MAR Simulation Conditions*

Parameter	N	Missing proportion											
		5%			10%			20%			30%		
		LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI
Coefficients	50	0.0996	0.0885	0.0860	0.1289	0.0901	0.1086	1.4705	0.0950	0.2284	5.4721	0.1053	0.3451
	100	0.0860	0.0778	0.0713	0.0979	0.0795	0.0853	0.1523	0.0812	0.1796	4.5217	0.0845	0.2935
	300	0.0740	0.0713	0.0618	0.0791	0.0721	0.0662	0.0970	0.0719	0.1703	0.1494	0.0733	0.2618
	500	0.0718	0.0704	0.0596	0.0743	0.0698	0.0612	0.0865	0.0709	0.1685	0.1179	0.0717	0.2616
Weights	50	0.2124	0.2056	0.1988	0.2422	0.2114	0.1926	0.5849	0.2319	0.1964	1.4292	0.2584	0.2134
	100	0.1971	0.1929	0.1841	0.2063	0.1955	0.1739	0.2750	0.2025	0.1588	0.7302	0.2132	0.1561
	300	0.1859	0.1855	0.1749	0.1885	0.1863	0.1597	0.2006	0.1880	0.1335	0.2630	0.1904	0.1218
	500	0.1839	0.1837	0.1727	0.1848	0.1842	0.1552	0.1916	0.1852	0.1276	0.2190	0.1868	0.1135

Note. This table displays the mean RMSE for pooled regression coefficients and pooled component weights across 1,000 iterations in all simulation conditions where the missing mechanism is **MAR**. RMSE = root mean squared error, MAR = missing at random, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.

Table C9*RMSE Results: MNAR Simulation Conditions*

Parameter	N	Missing proportion											
		5%			10%			20%			30%		
		LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI	LWD	MICE	LSI
Coefficients	50	0.1050	0.0912	0.0850	0.1277	0.0984	0.0933	0.1774	0.1120	0.1459	0.4974	0.1383	0.2496
	100	0.0932	0.0826	0.0712	0.1085	0.0859	0.0744	0.1368	0.0944	0.1134	0.1745	0.1137	0.2123
	300	0.0842	0.0752	0.0611	0.0958	0.0784	0.0580	0.1195	0.0881	0.0978	0.1397	0.1023	0.1851
	500	0.0820	0.0739	0.0595	0.0930	0.0768	0.0546	0.1134	0.0856	0.0982	0.1343	0.0999	0.1852
Weights	50	0.2004	0.2065	0.1961	0.2117	0.2107	0.1958	0.2616	0.2256	0.2153	0.4184	0.2685	0.2555
	100	0.1838	0.1905	0.1778	0.1841	0.1913	0.1707	0.1999	0.1926	0.1632	0.2407	0.2037	0.1745
	300	0.1721	0.1812	0.1686	0.1673	0.1783	0.1579	0.1654	0.1732	0.1411	0.1716	0.1708	0.1302
	500	0.1700	0.1792	0.1667	0.1644	0.1756	0.1550	0.1576	0.1688	0.1360	0.1590	0.1639	0.1233

Note. This table displays the mean RMSE for pooled regression coefficients and pooled component weights across 1,000 iterations in all simulation conditions where the missing mechanism is **MNAR**. RMSE = root mean squared error, MNAR = missing not at random, LWD = listwise deletion, MICE = multiple imputation by chained equations, LSI = least-squares imputation.