Stepped Wedge Cluster Randomized Controlled Trials for Three-Level Data: Design and Evaluation

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Ranran Dong, M.S.

Graduate Program in Biostatistics

The Ohio State University

2018

Dissertation Committee:

Abigail Shoben, Ph.D., Advisor

Rebecca Andridge, Ph.D.

Eloise Kaizar, Ph.D.

Michael Pennell, Ph.D.

Abstract

Cluster randomized trials (CRTs) are designed to randomly allocate groups of participants rather than individuals to either the intervention or the control. As an increasingly popular type of CRT, the stepped wedge CRT (SW-CRT) is a one-way crossover design where the intervention is provided sequentially to clusters whose orders are randomly determined.

In this dissertation, we propose novel SW-CRTs for three-level data such as patients (observation level) within wards (unit level) within hospitals (cluster level). The proposed designs differ in timing of allocating units within the same cluster to different treatments. We evaluate the efficiency of each design under a variety of underlying models generating three-level data. Impacts of misspecifying random unit effects and ignoring contamination on inference about the intervention effect are also evaluated via simulation studies.

We derive the closed-form expression for variance of the intervention effect estimator under a standard three-level model incorporating constant random unit effect across time. The formula is flexible and can accommodate a wide variety of three-level CRTs. Using the variance formula, we compute and compare the efficiencies across all of our proposed three-level SW-CRTs under various scenarios. Results show that when there is no contamination, design 4 transferring units from the same cluster at all time points is most efficient, and design 1 transferring units from the same cluster

at a single time point is least efficient. We then extend the standard three-level model by including varying random unit effects across time. Under the extended model, the order of efficiency among the proposed designs does not change.

For our proposed designs, we study the impact of model misspecification on inference about the intervention effect. We fit the standard model to data generated from the aforementioned extended model. Results show minimal influence on bias of the treatment effect estimator, but potentially low coverage probabilities for the treatment effect under all designs. Other studies we conduct include incorrectly assuming a random unit effect when none are present and incorrectly omitting a random unit effect when it truly exists. In these two cases, there is minimal impact on inference about the intervention effect under all designs.

We also address the problem of contamination for the proposed three-level SW-CRTs in which units from the same cluster may contaminate each other. Under each design, we consider practical scenarios where contamination could occur, describe the severity of contamination, and evaluate the impact of ignoring contamination when modeling in the presence of contamination. Our numerical studies show that designs 2 and 3 transferring units from the same cluster within two steps can still provide valid inference about the intervention effect in the presence of mild to moderate contamination, and design 4 should not be preferred even when contamination is mild.

Dedicated to my family.

Acknowledgments

My deepest and sincere gratitude goes to my advisor, Dr. Abigail Shoben. Thank you for providing priceless guidance on our research work, imparting your knowledge and wisdom to me without reservation, and teaching me to be a better collaborator. I will miss all the weekly meetings and fun discussions we had together over the past three years. I would also like to thank my candidacy and dissertation committee members, Dr. Rebecca Andridge, Dr. Eloise Kaizar, and Dr. Michael Pennell, for their constructive and insightful comments on my research.

My special thanks go to the Department of Statistics at The Ohio State University for providing me with financial support during my graduate study. My positive working experience in this department will have far-reaching impact on my future career. I would also like to thank former Vice Chair for Graduate Studies, Dr. Elizabeth Stasny, for fighting for me when I was applying to Ph.D. Program in Biostatistics at our university. Thank you for helping me win the opportunity to enter this great program.

Lastly, I would like to express my genuine love and gratitude to my family. I thank my parents, Yubing Dong and Yanli Yao, for their unconditional love and support, even though it may still be hard for them to accept that we are oceans apart. I would also like to thank my fiancé, Zhifei Yan, for his love, support, and help in my academic growth. I am more than lucky to have him on this graduate school journey.

Vita

2011 ..... B.S. in Statistics, Zhongnan University of Economics and Law, China.

2015 ..... M.S. in Statistics, The Ohio State University.

2013–2017 ..... Graduate Teaching Associate, The Ohio State University.

Fields of Study

Major Field: Biostatistics

Table of Contents

Appendices

List of Tables

List of Figures

Figure **Page**

Chapter 1: Introduction

Cluster randomized trials (CRTs) are designed to randomly allocate groups of participants rather than individuals to either the intervention or the control. They are particularly useful when it is impossible or inappropriate to randomize on the individual level (Bland, 2004). An important motivation of conducting CRTs is to avoid contamination which, in clinical trials, means that individuals in the control arm are exposed to the intervention although they are not supposed to, or vise versa. This contamination can happen when participants from different trial arms are close in geographical areas or able to communicate. For example, nurses receiving new training in the intervention arm may influence the behavior of nurses receiving standard training in the control arm in the same clinic by communication. To avoid contamination, an increasing number of studies recommend conducting randomization at cluster level rather than at individual-participant level. (Bland, 2004).

Among different cluster randomized trials, the stepped wedge cluster randomized trial (SW-CRT) is increasingly popular for its practical merits. The SW-CRT is essentially a one-way crossover design, where the intervention is provided sequentially to the clusters whose orders are randomly determined (Brown and Lilford, 2006). The stepped wedge design includes an initial phase where all of the clusters are assigned to the control arm. In the second phase of the study, some of the participating clusters

are randomly selected to transfer from the control arm to the intervention arm, while the remaining clusters still receive the control. Then in the third phase, another batch of clusters are randomly selected to transfer to the intervention, while clusters that finished the transition in the second phase continue receiving the intervention. The remaining clusters stay in the control arm. Following this procedure, all participating clusters eventually complete the transition from the control to the intervention. This feature serves as one of the most important advantages of stepped wedge designs, especially when the intervention is more likely to be beneficial than harmful (Brown and Lilford, 2006; Hemming et al., 2015a).

Traditional CRTs mainly deal with two-level data such as patients (observation level) from hospitals (cluster level) (Doig et al., 2008; Wood et al., 2008; Scales et al., 2016). Henceforth, we call them two-level CRTs. Over the past decade, practical applications about CRTs for three-level data, henceforth three-level CRTs, have been increasingly popular (Bruce et al., 2007; Marsteller et al., 2012; Juul et al., 2014). To give an example of three-level CRTs, we consider a cluster randomized controlled trial aiming at reducing the central line-associated bloodstream infection (Marsteller et al., 2012). The study took place in 45 intensive care units (ICUs) from 35 hospitals. The intervention were infection prevention practices that provided instructions to clinicians who inserted lines. A CRT was conducted at the hospital level to avoid contamination across different ICUs within the same hospital. Patient data such as whether or not an infection has occurred were collected. In this example, a three-level data structure was formed by patients (observation level), ICUs (unit level), and hospitals (cluster level). Other examples include nurses within general practices, wards within hospitals, and villages within geographical areas.

This chapter reviews design and analysis of standard cluster randomized trials, and problems of model misspecification and contamination in modeling data from CRTs. In Section 1.1, we introduce design and sample size calculations for the parallel CRT, the crossover CRT, and the stepped wedge CRT. Common methods for analyzing data collected from CRTs are also introduced. In Section 1.2, we summarize different directions of study on misspecifying random effects in linear mixed models and generalized linear mixed models. Section 1.3 provides important concepts and current work about contamination in clinical trials.

1.1 Design and Analysis of Cluster Randomized Trials

In this section, we first introduce common methods used for analyzing data collected from cluster randomized trials. Then we introduce standard designs in cluster randomized trials: the parallel CRT, the crossover CRT, and the stepped wedge CRT. For each standard CRT, we summarize its designs features, possible variations of the design, applications, and results about power and sample size calculations.

1.1.1 Analysis of Cluster Randomized Trials

In current work about estimating the treatment effect, there are mainly two classes of approaches: cluster-level analysis and model-based analysis. Cluster-level methods summarize data for each cluster into a single measure which in return is treated as raw data and used for later analysis (Matthews and Altman, 1990). For example, we can compute the mean response of each cluster that receives either the intervention or the control during each time period. Based on these cluster-level mean responses, we can construct a summary measure that estimates the intervention effect. Cluster-level analysis is simple and convenient. However, it does not allow for individual-level

covariates and thus ignores individual variations. Moreover, in a longitudinal study, cluster-level analysis may lead to bias in estimating the intervention effect if the response variable has trends in time (Turner et al., 2007).

On the other hand, model-based analysis methods provide more complex but potentially more efficient and unbiased estimators. These methods allow for relevant subject level covariates and time effect by including these sources of variation into the model. Common models used for clustered data analysis are the population average model which typically uses the generalized estimation equation (GEE) (Liang and Zeger, 1986) and the mixed model which usually uses likelihood-based approaches (Laird and Ware, 1982) for model fitting.

To compare between the mixed model and the population average model, we first consider an example which relates the length of walking per week to neighborhood crime rate using a generalized linear mixed model (GLMM) proposed as follows (Hubbard et al., 2010).

$$log(\frac{P(Y_{ij}=1|X_{ij},\alpha_j)}{1-P(Y_{ij}=1|X_{ij},\alpha_j)}) = \beta_0 + \alpha_{0j} + (\beta_1 + \alpha_{1j})X_{ij}, \tag{1.1}$$

where Y_{ij} is the response variable for the ith subject in the jth neighborhood. $Y_{ij} = 1$ if the subject walks more than 2 hours per week, and 0 otherwise. X_{ij} is a continuous measure of crime rate for neighborhood j which would be the same for each i within j. In addition, $\alpha_j = (\alpha_{0j}, \alpha_{1j})$ is the random effects for neighborhood j, and $\alpha_j \sim MVN(0, \Sigma)$, with Σ being a covariance matrix. Under this model, the interpretation of the nonintercept fixed effect is within the neighborhood level. Thus, given two neighborhoods having the same random effect, β_1 is the log odds ratio of walking for more than two hours/week comparing one neighborhood with crime rate one percent higher than the other neighborhood.

One pitfall of using the mixed model is that the misspecification of the joint distribution of the random effects and the error may be misleading under the likelihood estimation methods. Given model (1.1), consider the conditional density of Y_{ij} on X_{ij}:

$$L(Y_{ij}|X_{ij}) = \int_{\mathcal{R}} \int_{\mathcal{R}} f(Y_{ij}|X_{ij}, \alpha_{0j}, \alpha_{1j}) h(\alpha_{0j}, \alpha_{1j})(d\alpha_{0j})(d\alpha_{1j}). \tag{1.2}$$

There are an infinite number of combinations of f and h that can result in the same density L. The random-effects model for the density is nonidentifiable, since only the distribution of (Y_{ij}, X_{ij}) can provide information about the fit of competing models (Hubbard et al., 2010).

Under the population average model using generalized estimation equations (GEEs), the coefficients describe changes in the population mean when covariates changes. Unlike the mixed model, GEE does not require distributional assumptions, but only needs a few assumptions such as the mean of the conditional distribution of the response on covariates and the estimation function (Gardiner et al., 2009; Hubbard et al., 2010). Accordingly, the likelihood approach cannot be used for estimation. Robust or sandwich estimators have been proposed to estimate variance of the parameter estimators (Liang and Zeger, 1986). One advantage of these estimators is that even though the working covariance is misspecified, the asymptotic variance of the GEE estimator of model parameters is usually robust by using empirical estimators (Diggle et al., 2002). However, the number of independent groups needs to be sufficiently large so as to guarantee valid inference about parameter estimators. If one has a small number of large-sized clusters, extra assumptions may be added onto the working covariance (e.g., autoregressive, exchangeable, etc.) or more information in these large-sized clusters (Hubbard et al., 2010). In practice, the choice between

the marginal model (GEE) and the conditional model (mixed model) should depend on the relevant questions of interest (Gardiner et al., 2009).

Compared to cluster-level analysis methods, one possible disadvantage of model-based methods is that the derivation of statistical properties of the treatment effect estimator may be challenging, especially under complex designs.

1.1.2 Parallel CRT: Design, Power and Sample Size calculations

Design

Parallel CRTs have been extensively applied to many areas such as primary care and prevention studies (Simpson et al., 1995; Eldridgea et al., 2004; Pagoto et al., 2009; Simunovic et al., 2008), community level randomized trials (for example Ciliberto et al., 2005; Gruber et al., 2013), education (for example Murray et al., 1992; Bell et al., 1993; Kelly et al., 1991) and so on. Compared to other cluster designs, the parallel CRT is relatively convenient to implement and easy to understand. In addition, it does not require many repeated measurements on the same cluster, which puts less burden on care givers and participants.

Figure 1.1 shows two possible parallel designs, one with a baseline period and one without (Teenrenstra et al., 2012). Under the conventional parallel design (Figure 1.1(a)), clusters are assigned to either the intervention arm or the control arm. In our example, clusters of participants are followed over two time periods and stay in only one trial arm throughout the study. On the other hand, under the parallel design with a baseline period (Figure 1.1(b)), all six clusters are in the control arm during the baseline period. In the second time period, three clusters are randomly chosen to make have a crossover to the intervention arm, while the other three stay

in the control arm. An example of the parallel CRT with a baseline period can be found in the public policy evaluation applied to the Mexican universal health insurance program (King et al., 2007). In that study, all clusters were paired up so that one receive the intervention, while the other one received the control. The design with baseline measurements may yield more efficient parameter estimators than that without baseline measurements, since within-cluster comparisons are used in data analysis.

	Time	
Cluster	1	2
1	1	1
2	1	1
3	1	1
4	0	0
5	0	0
6	0	0

(a) Parallel CRT

	Time	
Cluster	1	2
1	0	0
2	0	0
3	0	0
4	0	1
5	0	1
6	0	1

(b) Parallel CRT with a baseline period

Figure 1.1: Examples of parallel CRT. In each cell, 0 means receiving the control, 1 means receiving the intervention, and "." means no data collected at the corresponding time point.

A natural extension of two-level parallel CRTs are three-level parallel CRTs where data contain the observation level, unit level and cluster level (Heo and Leon, 2008; Teerenstra et al., 2008). Figure 1.2 shows the design scheme of a three-level parallel CRT in which randomization is conducted at the cluster level. In this case, the design is essentially the same as the two level parallel CRT shown in Figure 1.1(a) except for the data structure.

Cluster	Unit	Time
		1
1	1	0
	2	0
	3	0
2	1	0
	2	0
	3	0
3	1	1
	2	1
	3	1
4	1	1
	2	1
	3	1

Figure 1.2: Three-level Parallel CRT. In each cell, 0 means receiving the control, 1 means receiving the intervention.

Power and Sample Size calculations

For the conventional two-arm parellel CRT with equal cluster sizes n and continuous responses, to reach the prespecified power $1-\beta$ given a sufficient number of clusters, the number of individuals m per arm is

$$\begin{aligned} m &= m_I \times [DE] \\ &= \frac{(z_{\alpha/2}+z_\beta)^2 2\sigma^2}{\Delta^2}[1+(n-1)\rho], \end{aligned} \tag{1.3}$$

where m_I is the required number of participants per arm for the individual randomized trial, σ^2 is the total variance of the response variable, and z_γ is the upper γth quantile of the standard Normal distribution (Donner et al. (1981); Shih (1997)). Assuming equal cluster size n, the required number of clusters per arm is m/n.

When the number clusters is fixed, Hemming et al. (2011) proposed formulae to determine the size of each cluster. For a fixed number k equally sized clusters in each

arm, the required cluster size n to reach power $1-\beta$ is

$$n = \lceil \frac{m_I(1-\rho)}{k - m_I\rho} \rceil,$$

where m_I is defined in (1.3). To make the above equation valid, we need the condition $k > m_I\rho$. If the condition is not satisfied (infeasible design), we could either determine the maximum available power to detect the pre-specified difference, or the minimum detectable difference under the pre-specified power. Once we find a feasible design, we can determine the required number of individuals n in each cluster using (1.1.2) again.

For binary responses with equal trial arms, Donner et al. (1981) obtain the following expression of the number of individuals m per arm

$$m = \frac{(z_{\alpha/2} + z_\beta)^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{\Delta^2}[1 + (n-1)\rho], \tag{1.4}$$

where π_1 and π_2 are the probabilities of the event of interest in the control arm and the intervention arm, respectively. The other notations are the same as in (1.3).

For count outcomes with equal trial arms, Amatya et al. (2013) use a GEE approach to determine the number of participants m per group given fixed cluster size

$$m = \frac{[z_{\alpha/2}\sqrt{2} + z_\beta\sqrt{1 + e^{-\tilde{b}}}]^2}{e^{\beta_0}\tilde{b}^2}[1 + (n-1)\rho],$$

where $\tilde{b}$ denotes the treatment effect, and β_0 denotes the event rate in the control arm.

For ordinal outcomes with equal trial arms, Kim et al. (2005) used the GEE containing ordinal repeated measurements to compute the required sample size via simulation studies. However, the performance of the method given a small number of large-size clusters is unknown. Alternatively, Campbell and Walters (2014) extended

the sample size calculation formula of Whitehead (1993) for ordered categorical data in individual randomized trials. They assumed a mixed model in which the treatment effect is evaluated by the log odds ratio. The required number of individuals m per arm is

$$m = \frac{6(z_{\alpha/2} + z_{\beta})^2/[log(OR)]^2}{1 - \sum_{i=1}^{I} \overline{\pi_i}^3}[1 + (n-1)\rho],$$

where $\overline{\pi_i}$ is the average expected proportion for the control and the intervention groups in ordinal category i.

Besides using the ICC to determine the sample size, Hayes and Bennet (1999) use the coefficient of variation (CV) and propose the sample size calculation formulae for both continuous and binary outcomes. For other types of outcome like time-to-event outcomes and rate outcomes, Rutterford et al. (2015) have provided a complete summary regarding the sample size calculations.

In practice, it is most likely that cluster sizes are unequal. In this case, Donner et al. (1981) computed the average cluster size to estimate the sample size needed:

$$DE = 1 + (\bar{n} - 1)\rho, \tag{1.5}$$

where $\bar{n}$ is the mean cluster size. However, Eldridge et al. (2006) argued that employing the average cluster size underestimates the true design effect. On the other hand, using the maximum cluster size may be too conservative. They therefore proposed the Maximum possible Inflation in sample Size (MIS) to estimate the required cluster sizes by including the coefficient of variation of cluster size which is defined as the standard deviation of the cluster size S_n divided by the average cluster size $\bar{n}$. This formula can be applied to both continuous and binary outcomes. In addition, they

concluded that when the coefficient of variation is less than 0.23, one may ignore the influence of adjustment for variable cluster size on sample size calculations.

Tokola et al. (2011) consider the test $H_0 : \Delta = 0$. Under Assumptions 1 (distributional assumptions) and 3 (design assumption) in the paper, they conclude that the distribution of the test statistic $T = \frac{\hat{\Delta}}{\hat{Var}(\hat{\Delta})}$ converges in distribution to a noncentral chi-square distribution given $H_1 : \Delta = \frac{\Delta_0}{\sqrt{N}}$, where N is the total number of subjects. Thus, the approximate power function of the level α test is

$$P\{\chi_1^2(\delta^2(N_1, n_1, N_0, n_0)) > \chi_{1,1-\alpha}^2\}, \tag{1.6}$$

where $\delta^2(N_1, n_1, N_0, n_0) = \frac{\Delta_1^2/\sigma^2}{\frac{1-\rho}{N_1}+\frac{\rho}{n_1}+\frac{1-\rho}{N_0}+\frac{\rho}{n_0}}$; σ^2 is the total variance of the response; N_0 and N_1 are the number of subjects in the control and the treatment arm, respectively; and n_0, n_1 are the number of clusters in the control and the treatment arm, respectively.

Under parallel designs with baseline measurements (Figure 1.1 (b)), study planners can carry out a cohort study or a cross-sectional study. In a cohort study, the same participants from each cluster are measured before and after the intervention is given, while in a cross-sectional design different participants from each cluster are measured before and after the intervention. It is widely accepted that cohort designs yield a theoretically more efficient estimator of the treatment effect than cross-sectional designs due to correlated responses being taken on the same participant over time. However, Feldman and McKinlay (1994) pointed out that some pitfalls of cohort designs such as follow-up cost and selection bias may outweigh its theoretical advantage, proposing a unifying model that includes both types of studies. Another problem of cohort designs is the aging of cohorts, which is more or less confounded with the change in the response. Even if the confounding is independent of the intervention assignment,

a change in the response variable in a cohort study cannot be directly compared with that in a cross-sectional study. Hence, a cross-sectional designs may be preferred if the question of interest is the difference in the response variable on the cluster level rather than on the individual level (Cambell et al. (2007)).

The relative estimation efficiency of cohort and cross-sectional designs is (Feldman and McKinlay (1994)):

$$\frac{Var_{cohort}(\hat{\Delta})}{Var_{cross-sectional}(\hat{\Delta})} = \frac{nM(1-\rho_c)\sigma_c^2 + M(1-\rho_s)\sigma_s^2 + \sigma_e^2}{nM(1-\rho_c^2)\sigma_c^2 + M\sigma_s^2 + \sigma_e^2},$$

where M is the number of replicate measurements per individual at each time point, ρ_c is the cluster autocorrelation which indicates the constant correlation among discrete time points, ρ_s is the subject autocorrelation which illustrates the correlation of a subject's responses among different time points, and σ_c^2, σ_s^2 and σ_e^2 are the variance components of the distribution of the random cluster effect, random subject effect and random error, respectively.

For the three-level parallel CRT shown in Figure 1.2, Heo and Leon (2008) derived an analytic form for power of the intervention effect based the Wald test under a three-level model with both cluster and random unit effects. Sample size calculations are then provided based on the power formula. However, their results are limited, since the three-level model they considered fails to include time effects. In this case, the GLS estimator of the intervention effect is simply the difference in mean responses between the intervention arm and the control arm. Under another three-level model for repeated measurements in parallel designs, Heo and Leon (2009) considered fixed intervention effect, fixed time effect, and fixed intervention-by-time interaction effect which is the parameter of interest. The interaction is estimated by taking the difference in the ML estimators of slope for the outcome between the intervention arm

and the treatment arm. Based on the Wald test and the variance of the estimator of the intervention-by time interaction, power and sample size calculations are then derived. Under this special problem of interest, the estimator of the intervention-by time interaction is again constructed from the mean response of each treatment arm. It is left unknown how to estimate the main effect of intervention under this more complex model with time effects.

1.1.3 Crossover CRT: Design, Power and Sample Size calculations

Design

While the development of parallel CRTs have been quite mature, the main challenge for the design is that the study groups should be balanced on relevant variables. By incorporating a cross-over in the randomized CRT, we can limit the influence of the imbalance of characteristics between the two arms, since all of the within-cluster comparisons are employed (Reich and Milstone, 2014). Furthermore, the cross-over CRT allows a limited number of clusters. Given the same number of participants, the crossover CRT yields higher power than the parallel CRT does. By controlling the within-cluster variation via the crossover CRT, we may obtain more efficient statistical comparisons than in the parallel CRT (Reich et al., 2012; Reich and Milstone, 2014).

Figure 1.3 displays the design scheme of a standard cross-over CRT. During time period 1, three clusters are randomized to the intervention arm and the other three are randomized to the control arm. In the next time period, clusters 1–3 that previously receive the intervention switch to the control arm, while clusters 4–6 that previously receive the control switch to the intervention arm.

	Time	
Cluster	1	2
1	1	0
2	1	0
3	1	0
4	0	1
5	0	1
6	0	1

Figure 1.3: Design scheme of standard cross-over CRT. In each cell, 0 means receiving the control, and 1 means receiving the intervention.

Although cross-over CRTs may reduce the impact of imbalance on relevant covariates and yield higher power than parallel CRTs, we still need to be cautious when applying them to practical problems. To begin with, cross-over studies eliminate confounding effects related to internal cluster features, but do not eliminate external factors such as a time effect. Therefore, the length of time periods should be shorter than anticipated changes in a cluster (Parienti and Kuss, 2007). Second, there exists potential risk of cluster-level carry-over effect in a cross-over design. Study planners need to choose an appropriate period length, or set a washout period whenever a cross-over happens (Parienti and Kuss, 2007).

Like the parallel CRT, the cross-over CRT has been extensively applied in practice. For example, to compare 10% povidone iodine aqueous solution with 5% PVP-I in a 70% ethanol-based aqueous solution for prevention of catheter colonization and catheter-related infection, researchers have conducted a cross-over CRT on two similar units of 11-bed adult medical intensive care in a France hospital (Parienti et al., 2004). The study lasted for 12 months, with every three months a cross-over happening

within each unit, as shown in Figure 1.4. In this example, each unit had three cross-overs and received both treatments for two time periods.

	Time			
Cluster	1	2	3	4
1	0	1	0	1
2	1	0	1	0

Figure 1.4: Design scheme of a more complex cross-over CRT. In each cell, 0 means using the 10% povidone PVP-I protocol, and 1 means using the new alcoholic PVP-I protocol.

Power and Sample Size calculations

Turner et al. (2007) compared several model-based and cluster-level approaches in the analysis of cluster randomized cross-over trial data. In the model based approaches, one model includes random cluster effects and the other model includes fixed cluster effects. Cluster-level analysis is based on within-cluster comparisons. They considered unweighted analysis, analysis weighted by cluster sizes, and analysis weighted by the combination of cluster sizes and estimated ICC. Via numerical examples, they compared these methods in empirical precision, coverage probabilities, and practical considerations. The conclusion is that the model-based analysis incorporating random cluster effects and the two weighted cluster-level analysis methods consistently perform well across all scenarios considered. Further, to choose between a model-based approach and a cluster-level analysis method, study planners also need to consider the extent of complexity in the analysis.

Giraudeau et al. (2008) proposed a within-cluster comparison approach for estimating the treatment effect estimator in the crossover CRT, assuming a cross-sectional

trial. Then derived the required number of clusters to reach power $1-\beta$

$$m = 2\frac{(z_{\alpha/2}+z_{\beta})^2\sigma_X^2\{\frac{1+(p_0-1)\rho}{p_0}-\eta\}}{(\mu^{(1)}-\mu^{(2)})^2},$$

where σ_X^2 is the variance component of the distribution of the outcome, p_0 is the number of participants in each arm from each cluster, ρ is the intra-cluster correlation, and $\mu^{(1)}$, $\mu^{(2)}$ are the mean responses of the two arms, respectively.

Unlike the parallel CRT, there has not been much theoretical work on power and sample size calculations for crossover CRTs. Most work focuses on simulation-based approaches to analyze the behavior of power and estimate required sample sizes. For example, Reich et al. (2012) used a generalized linear mixed model (GLMM) with Poisson outcomes to generate data under the cross-over CRT. They displayed a curve showing that the power for detecting the intervention effect increases as the number of independent clusters increases while holding the total number of participants constant. Given such curve, one may estimate the required number of clusters so as to achieve a pre-specified power level. In addition, based on the numerical example, one may obtain slightly higher power when the time effect is absent compared to the case when time-varying prevalence exists. An R package *clusterPower* has also been developed to calculate power for crossover CRTs using a simulation-based approach.

A few application studies also include power and sample size calculations for crossover cluster randomized designs. For example, in a study about chest radiographs for mechanically ventilated patients in intensive care units (ICUs) in 18 hospitals in France, researchers wanted to compare routine daily chest radiographs and an on-demand strategy in which chest radiographs are given only if warranted by the patient's clinical status (Hejblum et al., 2009). A crossover cluster randomized trial was conducted, with 21 independent ICUs divided into two arms having 10 and 11

units, respectively. The two trial arms used the routine or the on-demand strategy for chest radiographs during the first treatment period, then used the alternative strategy in the second period. As for sample size calculations, by simulation studies they decided to recruit 20 patients per strategy and a total of 800 patients so as to detect substantial differences in mortality or mean duration of mechanical ventilation. With the type I error controlled at 5% and the power reached at 80%, the calculated sample size should be able to detect a difference of 10% in mortality rate and a difference of 3 days in mean duration of mechanical ventilation between the two trial arms.

1.1.4 Stepped Wedge CRT: Design, Power and Sample Size calculations

Design

Figure 1.5 displays the design scheme of a complete SW-CRT. The intervention is rolled out sequentially to 6 clusters, each of which follow a different design pattern. In the initial stage, all clusters are assigned to the control arm. During time period 2, cluster 1 is selected to have a crossover to the intervention arm, while the other five clusters remain in the control arm. Furthermore, cluster 1 stays in the intervention arm towards the end of the study. In time period 3, cluster 2 switches to the intervention arm, while clusters 3–6 remain in the control arm. Similar to cluster 1, cluster 2 stays in the intervention arm towards the last time period of the study. Following this pattern, all clusters are eventually exposed to the intervention at the end of the study. The "steps" in the design scheme are formed by clusters having their one-way crossovers at different times.

	Time						
Pattern	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	0	1	1	1	1	1	1
2	0	0	1	1	1	1	1
3	0	0	0	1	1	1	1
4	0	0	0	0	1	1	1
5	0	0	0	0	0	1	1
6	0	0	0	0	0	0	1

Figure 1.5: The stepped wedge design for two-level data.In each cell, 0 means receiving the control, and 1 means receiving the intervention.

The earliest known application of the SW-CRT can be found in a hepatitis B (HBV) study (Gambia Hepatitis Study Group, 1987). The researchers in the study conducted a phased randomized trial so as to introduce the HBV vaccination to infants in Gambia. In this study, vaccination teams were treated as independent clusters according to geographical areas. There were 17 vaccination teams in total, in which a single team was randomized at each step of the trial. The duration of each step was approximately 10 to 12 weeks such that complete national coverage of the HBV vaccination should have been obtained within about 4 years. More recent applications of the SW-CRT include HIV studies [e.g., Hughes et al. (2003); Killama et al. (2010), patient safety intervention [e.g., Brown et al. (2008), children's malnutrition problems [e.g., Ciliberto et al. (2005) and many others. Two review articles systematically describe the applications of the SW-CRT to various research domains (Brown and Lilford, 2006; Mdege et al., 2011).

Stepped wedge designs have been increasingly employed for several reasons. First, it may be unethical to withhold an intervention in a parallel design or withdraw an intervention in a crossover design if previous evidence shows effectiveness of the intervention (Brown and Lilford, 2006; Hemming et al., 2015a). In a stepped wedge

design, every cluster will eventually receive the intervention. Second, the stepped wedge design is particularly useful when financial or logistical constraints require the intervention to be rolled out in stages (Cook and Campbell, 1979). For example, consider that a medical team travels village by village in some country in Africa to give HIV treatments to infected residents. It may not be feasible that the team provides their service to two or more villages simultaneously due to labor constraints.

Besides the possible advantages, some potential pitfalls of the stepped wedge design should also be considered in practice. Multiple stages in a stepped wedge design may lead to longer trial duration and higher data collection costs compared to many other designs (Brown and Lilford, 2006; Hussey and Hughes, 2007; Brown et al., 2008). Furthermore, a stepped wedge design may place a heavy burden on both participants and researchers, especially in a cohort design where each participant needs to be followed for multiple time periods. Accordingly, the quality of data collection may be undermined (Lynn, 2009).

In addition to the conventional SW-CRT, several variations for the stepped wedge design has been prposed by Hemming et al. (2015b). First, instead of conducting a complete SW-CRT where all clusters are being measured throughout the study, an incomplete trial can be conducted to reduce the data-collecting burden on researchers and participants (Figure 1.6 (a)). Under this design, each of the five clusters is measured one time before the exposure to the intervention and two times after the intervention. Second, there may be an implementation or transition phase of each cluster when it is transferring from the control arm to the intervention arm as in Figure 1.6 (b). Clusters during the transition periods can neither be considered in the control arm nor fully exposed to the intervention. No data are collected on a

cluster during its implementation phase. The third design is essentially a variation on the parallel CRT where the randomization is staggered while the design is balanced on time. In Figure 1.6 (c), at each of the three time points, two clusters are randomly selected to be randomized to either the treatment or control. Although in this design not all clusters complete the transition from the control to the intervention as in an SW-CRT, it is pointed out that staggered parallel CRTs can also be considered under the same framework as the stepped wedge design. Lastly, the three-level SW-CRT is proposed for data that contain two layers of clustering. Figure 1.6 (d) provides the design scheme for the three-level SW-CRT. In this case, the difference between the three-level and the two-level SW-CRTs is minimal except for how observations are clustered.

	Time						
Cluster	1	2	3	4	5	6	7
1	0	1	1	.	.	.	.
2	.	0	1	1	.	.	.
3	.	.	0	1	1	.	.
4	.	.	.	0	1	1	.
5	.	.	.	.	0	1	1

(a) An incomplete SW-CRT with one before and two after measurements

	Time						
Cluster	1	2	3	4	5	6	7
1	0	.	1	1	1	1	1
2	0	0	.	1	1	1	1
3	0	0	0	.	1	1	1
4	0	0	0	0	.	1	1
5	0	0	0	0	.	1	1
6	0	0	0	0	0	.	1

(b) An incomplete SW-CRT with an implementation period

	Time							
Cluster	1	2	3	4	5	6	7	8
1	0	.	.	.	.	0	.	.
2	0	.	.	.	.	1	.	.
3	.	0	.	.	.	.	0	.
4	.	0	.	.	.	.	1	.
5	.	.	0	.	.	.	.	0
6	.	.	0	.	.	.	.	1

(c) Staggered parallel CRT with baseline measurements

Cluster	**Unit**	**Time**				
		1	**2**	**3**	**4**	**5**
1	**1**	0	1	1	1	1
	2	0	1	1	1	1
	3	0	1	1	1	1
2	**1**	0	0	1	1	1
	2	0	0	1	1	1
	3	0	0	1	1	1
3	**1**	0	0	0	1	1
	2	0	0	0	1	1
	3	0	0	0	1	1
4	**1**	0	0	0	0	1
	2	0	0	0	0	1
	3	0	0	0	0	1

(d) Three-level SW-CRT

Figure 1.6: Variations on SW-CRTs. In each cell, 0 means receiving the control, 1 means receiving the intervention, and "." means no data are collected at the corresponding time point.

Power and sample size calculations

Consider the following standard two-level model for SW-CRTs with continuous outcomes (Hussey and Hughes, 2007).

$$Y_{ijk} = \mu + a_i + \delta_j + X_{ij}\theta + \epsilon_{ijk}, \tag{1.7}$$

where $i = 1, \ldots, I$, I being the number of clusters, $j = 1, \ldots, T$, T being the number of time points, and $k = 1, \ldots, N$, N being the number of individuals measured in each cluster at each step. Further, a_i is the random cluster effect, δ_j is the fixed time effect, and θ is the fixed treatment effect. The model assumes $a_i \sim N(0, \tau^2)$, $\epsilon_{ijk} \sim N(0, \sigma_e^2)$, and a_i is independent of ϵ_{ijk}. X_{ij} is the indicator variable for the treatment (1 if receiving treatment, and 0 otherwise). The intra-cluster correlation (ICC) is defined to be $\rho = \frac{\tau^2}{\tau^2+\sigma_e^2}$, which is the correlation among individuals in the same cluster.

Based on model (1.7), Hussey and Hughes (2007) derived the analytic form of power for the test $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$. Test statistic $Z = \frac{\hat{\theta}-\theta_0}{\sqrt{\text{Var}(\hat{\theta})}}$ is used. As a key component in the derivation of power formula, the variance of the GLS estimator of the intervention effect $\hat{\theta}$ is given by

$$\text{Var}(\hat{\theta}) = \frac{I\sigma^2(\sigma^2 + T\tau^2)}{(IU - W)\sigma^2 + (U^2 + ITU - TW - IV)\tau^2}, \tag{1.8}$$

where $U = \sum_{i,j} X_{ij}$, $W = \sum_j (\sum_i X_{ij})^2$, $V = \sum_i (\sum_j X_{ij})^2$, and $\sigma^2 = \sigma_e^2/N$. In practice, investigators are more interested how the ICC ρ affects the variance of the intervention effect estimator. Thus, expressed in terms of ρ, the above variance formula becomes

$$\text{Var}(\hat{\theta}) = \frac{I\sigma_e^2[1 + (NT - 1)\rho]}{(IU - W)N(1 - \rho) + (U^2 + ITU - TW - IV)N^2\rho}.$$

In addition, we take note that the treatment indicator X_{ij} is general, and the above variance formula works for any types of CRTs with two-level data under model (1.7).

Based on (1.8), Woertman et al. (2013) derived the design effect (DE) for the stepped wedge design based on the relative efficiency of the parallel individual randomized trial to the SW-CRT given the same number of individual observations $NI(b+tk)$. In the two-arm parallel individual randomized trial, the variance of the intervention effect estimator $\widehat{\theta_z}$ based on the two-sample Z-test is $Var(\widehat{\theta_z}) = \frac{4\sigma_y^2}{NI(b+tk)}$. Thus, the design effect for the SW-CRT is

$$DE = \frac{Var(\hat{\theta})}{Var(\widehat{\theta_z})} = \frac{1+\rho(ktN+bN-1)}{1+\rho(\frac{1}{2}ktN+bN-1)} \frac{6(1-\rho)}{t(k-\frac{1}{k})} \frac{\sigma_y^2}{Nik} / \frac{4\sigma_y^2}{NI(b+tk)} \tag{1.9}$$

where $k = T-1$ is the number of steps, t is the number of measurements after each step, b is the number of baseline measurements, and i is the number of clusters making the switch from the control to the intervention at each step.

Simulation-based methods have also been used for power and sample size calculations to satisfy more special and complex underlying models for SW-CRTs (Dimairo et al., 2011; Baio et al., 2015). For example, Baio et al. (2015) compared sample size requirements for the SW-CRT and the parallel CRT via simulation studies. Both continuous and discrete outcomes are modeled under both cross-sectional design and cohort design. Besides the standard model (1.7), another model they considered is the standard model with added cluster-specific treatment random effect to account for different treatment effects on different clusters. For each model and type of outcomes, they studied the relations of power and the ICC and discussed sample size calculations under both parallel and stepped wedge CRTs.

1.2 Model Misspecification of Random Effects

One interesting question about model misspecification in mixed effects models is the misspecification of random effects. As pointed out by Hubbard et al. (2010), there are numerous possible choices for the distribution of random effects in a mixed model leading to the same marginal specification. When random effects are misspecified in a model, the interpretation of the fixed effects of the model is still unclear even if they are correctly specified. In general, current work on consequences of misspecification of random effects in mixed models can be divided into two directions: the choice of distributions or shapes of random effects (Verbeke and Lesaffre, 1996, 1997; Heagerty and Kurland, 2001; Agresti et al., 2004; Litiere et al., 2007; Huang, 2009; McCulloch and Neuhaus, 2011; White, 1982) and the choice of covariance patterns of random effects (Kwok et al., 2007; Ferron et al., 2002; Liu et al., 2012).

In the first category, the popular normality assumption of the random effects in LMM or GLMM has been questioned and possible diagnostic methods have been provided. In Verbeke and Lesaffre (1996), an underlying LMM with random effects being a finite mixture of g normal distributions is considered. They generated data using the underlying model and fitted the data with an LMM with the same fixed effects but normal random effects. Their results show that the estimation of the fixed effects, covariance matrix and the residual variance are not largely influenced. Similarly, Verbeke and Lesaffre (1997) showed that misspecification of the distribution of random effects in LMMs may hardly impact parameter estimation of fixed effects. However, they also concluded that asymptotic properties of parameter estimators for fixed effects may no longer hold when the inverse Fisher information matrix is used to estimate their standard errors. Thus, valid inference about model parameters requires

a corrected version of standard errors rather than the inverse Fisher information matrix.

Other work has focused on random effects misspecification under the GLMM. Heagerty and Kurland (2001) investigated the asymptotic relative bias when the random effects are misspecified in logistic mixed models. They concluded that severe bias can result in the ML estimator of fixed effects when the distribution of random effects depends on covariates. Via simulation studies, Agresti et al. (2004) studied the fit of two logit models assuming either normal random effects or using a nonparametric fitting method, while the true distribution of the random effects is a two-point mixture with a large variance. Their results show large bias and potentially dropped efficiency in parameter estimation when the models are misspecified. Litiere et al. (2007) concluded that the Type I error rate can be largely inflated when the distribution of the random intercept in a logistic random-intercept model is misspecified. Moreover, depending on the shape of the underlying distribution of the random intercept, power can either be inflated or deflated.

Despite previous concerns about consequences of misspecifying distributions of random effects in mixed models, McCulloch and Neuhaus (2011) argued that these concerns may be misplaced. Reasons include that some previously studied situations are unfairly extreme, and that sensitivity to misspecification had not been adequately studied. Through their simulation studies based on logistic mixed models, they concluded that a wide range of inferences is quite robust to misspecification of random effects distributions, particularly for the estimation of within-cluster covariate effects which are of great interest in longitudinal studies.

The second direction of studies is on misspecification of covarianace patterns of random effects. Ferron et al. (2002) considered the effects of overly simplifying the covariance pattern of first-level random errors under multi-level model. The model is proposed as follows. At the first level,

$$Y_{it} = \pi_{i0} + \pi_{i1} a_{it} + \epsilon_{it},$$

where Y_{it} is the tth outcome for subject i, a_{it} is the time at which tth outcome on the ith subject is taken, and ϵ_{it} is the first-level random error term. At the second level,

$$\pi_{0i} = \beta_{00} + \beta_{01} x_{i1} + r_{0i},$$

$$\pi_{1i} = \beta_{10} + \beta_{11} x_{i1} + r_{1i}.$$

where x_{i1} is the predictor for the ith subject, β_{01} and β_{11} are respectively coefficients of x_{i1} for π_{0i} and π_{1i}, and r_{0i} and r_{1i} are random error terms. In their numerical study, the underlying covariance structure Σ of ϵ_{it} is assumed to be first-order autoregressive, and the assumed structure is simply a diagonal matrix ($\sigma^2 I$). Results show unbiased estimates of the fixed effects $\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11}$, but their estimated variances are systematically inflated. Kwok et al. (2007) provided more complete results and systematically investigated the effects of different forms of misspecification of the covariance matrix Σ of within-subject random errors under multi-level models for longitudinal data. They concluded that both over-simplified and general misspecification of Σ would lead to inflated estimation of variances of both fixed and random effects.

1.3 Contamination

In clinical trials, contamination is defined to be the process where participants in the control (or intervention) arm unexpectedly receive the intervention (or control). In most cases, contamination is referred in particular to the situation of control subjects being exposed to at least some of the intervention. For example, an intervention is designed to engage at-risk students with school and reduce dropout rates. At-risk students who are given the intervention may interact with other at-risk students who are not given the intervention within the same school. In this case, the at-risk students in the control arm may also increase engagement with school due to their peers in the intervention arm (Rhoads, 2011).

The most serious consequence of contamination is introducing bias in estimation of the treatment effect. By making responses in the control arm more similar to responses in the intervention arm, contamination dilutes the treatment effect and causes underestimation of the treatment effect. As a result, the power of detecting the intervention effect is reduced (Torgerson, 2001; Keogh-Brown et al., 2007).

In practice, cluster randomized trials are often used to avoid potential contamination. The ability of cluster randomized trials to minimize contamination by physically separating participants in different trial arms is the main reason for preferring cluster randomized designs to individual randomized trials (Rhoads, 2011). In the example of decreasing dropout rate, the intervention could be delivered at the school level so that all at-risk students within the same school are in the same trial arm. In this case, contamination within the same school is avoided. However, disadvantages of CRTs may be recruitment bias and reduced statistical efficiency. To alleviate these

problems, Borm et al. (2005) proposed pseudo cluster randomized trials carrying features of both cluster randomized designs and individual randomized designs . Pseudo cluster randomization is done in a two-step procedure. First, all participating clusters are divided into two groups. Second, within each group, individual participants are randomly allocated to either the intervention or the control. The majority of participants in one group need to receive the intervention, and the majority of participants in the other group need to receive the control. They concluded that pseudo cluster randomized trials are more efficient than CRTs and individual randomized trials when contamination is present.

1.4 Outline of the Dissertation

This dissertation focuses on the design and analysis of novel stepped wedge cluster randomized trials for three-level data such as patients (observations) within wards (units) within hospitals (clusters). Under a standard three-level model, we develop flexible analytical results for power and sample size calculations that can be applied to arbitrary CRTs. We propose a set of novel three-level SW-CRTs and study the efficiencies of the proposed three-level SW-CRTs using the developed analytical results under the standard three-level model. We also extend the standard model to more complex underlying models to generate data under each proposed design, and evaluate consequences of misspecifying random effects and omitting contamination when modeling data from three-level SW-CRTs.

In Chapter 2, we develop analytical results for power and sample size calculations for three-level CRTs under a standard three-level model including random cluster effects, random unit effects, and fixed time effects. Our results are flexible and can

be applied to arbitrary types of three-level CRTs. We also review special cases of our results on power calculations for both two-level and three-level CRTs.

In Chapter 3, we propose a set of SW-CRTs for three-level data. The proposed SW-CRTs differ in timing of allocating different units from the same cluster to the intervention. We give four examples of our proposed designs and provide their design schemes along with practical usefulness. Using the developed analytical results from Chapter 2, we compare the efficiencies of the four designs using variance of the intervention effect estimator under the standard model introduced in Chapter 2. Illustrating patterns in variance of the intervention effect estimator as within-cluster correlations change, we also identify three different types of comparisons between the treatment group and the control group for estimating the treatment effect.

In Chapter 4, we extend the standard three-level model to one with varying random unit effects across time points. We choose a Toeplitz covariance pattern as the underlying covariance pattern of the random unit effects across time points. When the underlying model is correctly specified, we study how variance of the treatment effect estimator change with varying within-cluster correlations. In addition, we study the consequences of incorrectly specifying random unit effects to be constant across all times, while data is generated from the extended model with varying random unit effects across time points.

In Chapter 5, we study the consequences of omitting contamination when control units are contaminated by intervention units within the same cluster under our proposed designs. For each design, we illustrate practical scenarios where contamination could occur, describe severity of contamination, and evaluate the impact of ignoring contamination when modeling data in the presence of contamination. We consider

alternative models to generate data from three-level SW-CRTs with or without contamination. In addition, we provide simulation studies to evaluate consequences of wrongly adding or omitting random effects in these alternative three-level models.

Chapter 6 provides a summary of this dissertation and some future directions of our research. We will extend our proposed designs to incomplete designs where no data collection on some participants is needed during certain time periods. We also include alternative underlying models generating data from SW-CRTs, and discuss cohort designs that also fit into the framework of SW-CRTs.

Chapter 2: A General Formula for Power and Sample Size Calculations for Arbitrary Cluster Randomized Trials

In this chapter, we focus on power and sample size calculations for cluster randomized controlled trials for three-level data. We derive the closed-form expression for power of testing the intervention effect under a standard three-level model incorporating time effects, as most trials last a long time and patients outcome may change over time. Our power formula is flexible and can be applied to a wide variety of three-level CRTs. The results also include previous related results as special cases. We show that the derived power formula for testing the intervention effect based on its GLS estimator under our standard three-level model can be simplified to the power formula under the standard two-level model in Hussey and Hughes (2007). Also, the three-level model that we consider includes the model for the three-level parallel CRT considered in Heo and Leon (2008) as a special case. We show that their power formula can be directly recovered by our result in the special case of the parallel design with a single time period.

The rest of this chapter is organized as follows. In Section 2.1, we introduce a standard three-level model incorporating time effects. In Section 2.2, we derive the closed-form expression for the variance of the GLS estimator of the intervention effect under the three-level model and provide the power formula under arbitrary designs for

three-level data. We also review special cases of our results about power calculations for three-level CRTs. Sample size calculations are then introduced. In Section 2.3, we illustrate possible applications of our power formula to sample size calculations using a real-world example. Our conclusion appears in Section 2.4.

The following notations are used hereafter. Let $\mathbf{0}_n$ and $\mathbf{1}_n$ denote $n \times 1$ vectors of all 0s and all 1s, respectively. Let I_n be the $n \times n$ identity matrix and J_n be the $n \times n$ matrix of all 1s. Use $\otimes$ to denote the Kronecker product of matrices. Using Kronecker product, we express a block diagonal matrix with all m diagonal blocks being the same matrix A as $I_m \otimes A$.

2.1 The Model

Let us consider a linear mixed effects model that describes data collected from an arbitrary three-level cluster randomized controlled trial

$$Y_{ijtk} = \delta_t + Z_{ijt}\theta + a_i + b_{ij} + \epsilon_{ijtk}, \tag{2.1}$$

where $i = 1, \cdots, I$ is the index for independent clusters, $j = 1, \cdots, J$ is the index for units nested within the same cluster, $t = 1, \cdots, T$ is the index for time points, and $k = 1, \cdots, K$ is the index for observations from the same unit within a cluster at the same time. Accordingly, Y_{ijtk} denotes the kth observation from the jth unit nested in the ith cluster at time t. Z_{ijt} is the binary intervention indicator for the jth unit within the ith cluster at time t. Z_{ijt} takes value of 1 if the intervention is received, and 0 if the control is received. The fixed time effect at time t is denoted by δ_t. The intervention effect is denoted by θ. The cluster-level random effects $a_i, i = 1, \cdots, I$ are independently drawn from $N(0, \sigma_a^2)$. Within the same cluster i, the random unit effects $b_{ij}, j = 1, \cdots, J$ are independently drawn from $N(0, \sigma_b^2)$. The random errors

ϵ_{ijtk} are independently drawn from $N(0, \sigma_e^2)$. Furthermore, the random effects a_i, b_{ij}, and the error ϵ_{ijtk} are mutually independent. A similar model has been proposed by Hemming et al. (2015b) under a stepped wedge design. However, we note that our work is not limited to stepped wedge designs. Throughout this chapter, we focus on cross-sectional studies where the observations Y_{ijtk} are collected from $IJTK$ different participants.

One feature of model (2.1) worth noting is that it allows randomization to be at the unit level, not merely at the cluster level. Specifically, Z_{ijt} gives the assignment of the intervention to the jth unit within the ith cluster at time t. For participants from different units within the same cluster, their assignments of the intervention or the control are allowed to be different. In other words, model (2.1) allows that $Z_{ijt} \neq Z_{ij't}, \forall j \neq j'$.

2.2 Derivation of the flexible power formula for three-level CRTs

In this section, we derive the closed-form expression for power of testing the intervention effect under our three-level model (2.1). The test is based on the GLS estimator of the intervention effect. The main difficulty lies in deriving the variance formula for the GLS estimator of the intervention effect.

Without loss of generality, we take the cell mean over the K observations for ease of derivation. This yields the following cell-mean model

$$\overline{Y}_{ijt\cdot} = \delta_t + Z_{ijt}\theta + a_i + b_{ij} + \bar{\epsilon}_{ijt\cdot}, \tag{2.2}$$

where $\overline{Y}_{ijt\cdot} \coloneqq \sum_{k=1}^{K} Y_{ijtk}/K$, $\bar{\epsilon}_{ijt\cdot} \coloneqq \sum_{k=1}^{K} \epsilon_{ijtk}/K$. Thus, $\bar{\epsilon}_{ijt\cdot}$ is normal with mean 0 and variance $\sigma^2 \coloneqq \sigma_e^2/K$.

2.2.1 Preliminaries

Let us consider the following general normal linear model in matrix form

$$y = X\beta + \varepsilon, \tag{2.3}$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta = (\beta_1, \cdots, \beta_p)^T \in \mathbb{R}^p$, $\varepsilon \sim N_n(0, V)$, and $V \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix. We start with deriving a general formula for the variance of a single component in the GLS estimator $\hat{\beta}$ of β under model (2.3). Without loss of generality, let the single component of interest be the last entry β_p of β, since one can always permute the the component of interest to the last entry of β and permute the columns of X accordingly.

Theorem 1. *The variance of the GLS estimator $\hat{\beta}_p$ of β_p in (2.3) is*

$$\mathrm{Var}(\hat{\beta}_p) = \frac{1}{x_2^T V^{-1} x_2 - x_2^T V^{-1} X_1 (X_1^T V^{-1} X_1)^{-1} X_1^T V^{-1} x_2},$$

where $X = (X_1, x_2)$ is partitioned into the first $p-1$ columns $X_1 \in \mathbb{R}^{n\times(p-1)}$ and the last column $x_2 \in \mathbb{R}^n$.

The proof is provided in Appendix A.

2.2.2 Variance of the intervention effect estimator under three-level CRTs

In order to apply Theorem 1 when deriving the variance of the GLS estimator of the intervention effect $\hat{\theta}$, we write the cell-mean model (2.2) in matrix form. Due to the normality of the random effects and the error terms and their mutual independence, the cell-mean model (2.2) can be written as a general normal linear model taking the

exact form of (2.3), where

$$\begin{aligned} y \coloneqq (&\overline{Y}_{111\cdot}, \cdots, \overline{Y}_{11T\cdot}, \cdots, \overline{Y}_{1J1\cdot}, \cdots, \overline{Y}_{1JT\cdot}, \\ &\overline{Y}_{211\cdot}, \cdots, \overline{Y}_{21T\cdot}, \cdots, \overline{Y}_{2J1\cdot}, \cdots, \overline{Y}_{2JT\cdot}, \\ &\cdots, \overline{Y}_{I11\cdot}, \cdots, \overline{Y}_{I1T\cdot}, \cdots, \overline{Y}_{IJ1\cdot}, \cdots, \overline{Y}_{IJT\cdot})^T \in \mathbb{R}^{IJT}, \\ \beta \coloneqq (&\delta_1, \delta_2, \cdots, \delta_T, \theta)^T \in \mathbb{R}^{T+1}. \end{aligned}$$

In addition, we partition the design matrix X into (X_1, z), where X_1 and z are formed by the first T columns and the last column of X, respectively. Specifically,

$$\begin{aligned} X_1 \coloneqq\ & \mathbf{1}_{IJ} \otimes I_T, \\ z \ \coloneqq (&Z_{111}, \cdots, Z_{11T}, \cdots, Z_{1J1}, \cdots, Z_{1JT}, \\ &Z_{211}, \cdots, Z_{21T}, \cdots, Z_{2J1}, \cdots, Z_{2JT}, \\ &\cdots, Z_{I11}, \cdots, Z_{I1T}, \cdots, Z_{IJ1}, \cdots, Z_{IJT})^T \in \mathbb{R}^{IJT}. \end{aligned}$$

Under our model, X_1 is the design for the time effect and z is the vector of the treatment indicators. The error vector $\varepsilon \sim N_{IJT}(0, V)$, where $V \coloneqq I_I \otimes W$ and $W \coloneqq I_J \otimes (\sigma^2 I_T + \sigma_b^2 J_T) + \sigma_a^2 J_{JT}$.

Using Theorem 1, we can derive the closed-form expression for the variance of the GLS estimator of the intervention effect, $\mathrm{Var}(\hat{\theta})$.

Theorem 2. *The variance of the GLS estimator of θ in model (2.1) can be expressed in terms of the vector of the treatment indicators z:*

$$\begin{aligned} \mathrm{Var}(\hat{\theta}) &= \frac{1}{f(z)}, \ \textit{where} \\ f(z) \coloneqq\ & p \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T} Z_{ijt}^2 - pq \sum_{i=1}^{I}\sum_{j=1}^{J}(\sum_{t=1}^{T} Z_{ijt})^2 - r \sum_{i=1}^{I}(\sum_{j=1}^{J}\sum_{t=1}^{T} Z_{ijt})^2 \\ & - \frac{p}{IJ}\sum_{t=1}^{T}(\sum_{i=1}^{I}\sum_{j=1}^{J} Z_{ijt})^2 + \frac{pq + rJ}{IJ}(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T} Z_{ijt})^2, \end{aligned} \tag{2.4}$$

and $p \coloneqq \frac{1}{\sigma^2}, q \coloneqq \frac{\sigma_b^2}{\sigma^2+T\sigma_b^2}, r \coloneqq \frac{\sigma_a^2}{(\sigma^2+T\sigma_b^2+JT\sigma_a^2)(\sigma^2+T\sigma_b^2)}$.

Proof Sketch. According to Theorem 1, we have

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{z^T V^{-1} z - z^T V^{-1} X_1 (X_1^T V^{-1} X_1)^{-1} X_1^T V^{-1} z},$$

where $X_1 = \mathbf{1}_{IJ} \otimes I_T$, $V = I_I \otimes W$, and $W = I_J \otimes (\sigma^2 I_T + \sigma_b^2 J_T) + \sigma_a^2 J_{JT}$. The claimed result follows by deriving the closed-form expressions for V^{-1} and $(X_1^T V^{-1} X_1)^{-1}$. We leave the detailed proof to Appendix A. □

The variance formula for the GLS estimator $\hat{\theta}$ presented in Theorem 2 is flexible to design. The vector of treatment indicators z can accommodate a wide variety of designs for three-level data. This includes not only standard designs, but also hybrid designs that may carry features of standard designs.

Our results also include several previous results of power and sample size calculations for CRTs as special cases.

- The formula for $\mathrm{Var}(\hat{\theta})$ in Theorem 2 under the standard three-level model (2.1) can be simplified to the variance formula for the GLS estimator of the intervention effect under the standard two-level model in Hussey and Hughes (2007). In fact, if we let the random unit effect $b_{ij} = 0$, the number of units $J = 1$, and remove the redundant unit index j, then a cluster is equivalent to a unit. In this case, our model (2.1) simplifies to a two-level model with I independent clusters, within which there are K observations collected at each of the T time points, which gives exactly the same two-level model (1.7) considered in Hussey and Hughes (2007). Accordingly, as for the form of $\mathrm{Var}(\hat{\theta})$, in (2.4) we set $\sigma_b = 0, J = 1$, and drop the unit index j as well as any summation over j. This gives the variance formula (1.8) for arbitrary two-level CRTs.

- Our formula can be simplified to be the variance formula for the GLS estimator of the intervention effect under the three-level model for the parallel CRT with a single time period in Heo and Leon (2008). Our formula for $\mathrm{Var}(\hat{\theta})$ accounts for the time effect, since our model incorporates the fixed time effect that leads to the first block X_1 in the design matrix representing an one-way ANOVA design. If we consider the balanced parallel design with a single time period that assigns half of the clusters to the intervention arm and the other half to the control arm assuming that the number of clusters is an even number, then model (2.1) describes the three-level parallel CRT discussed in Heo and Leon (2008). In this case, X_1 reduces to an all-1s vector corresponding to the fixed intercept, and we can recover their result about the variance of the GLS estimator of θ from our formula (2.4) by setting $T = 1$, dropping the time index t in Z_{ijt} as well as any summation over t, and letting the vector of treatment indicator $z = (\mathbf{1}_{IJ/2}^T, \mathbf{0}_{IJ/2}^T)^T$. The resulting formula is

$$\mathrm{Var}(\hat{\theta}) = \frac{4}{IJK}\{\sigma_y^2 + K(J-1)\sigma_a^2 + (K-1)(\sigma_a^2 + \sigma_b^2)\} \tag{2.5}$$

(equation (11) in Heo and Leon, 2008).

2.2.3 Power and sample size calculations

To perform the following test of the intervention effect $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$, the test statistic $Z = \frac{\hat{\theta}-\theta_0}{\sqrt{\mathrm{Var}(\hat{\theta})}}$ is used when the variance components σ_a^2, σ_b^2 and σ_e^2 are known. For any alternative θ_a, the approximate power of the test at significance level α is

$$\mathbb{P}_{\theta_a}(\text{Reject H}_0) = \Phi(\frac{|\theta_a - \theta_0|}{\sqrt{\mathrm{Var}(\hat{\theta})}} - z_{\alpha/2}), \tag{2.6}$$

where $\text{Var}(\hat{\theta})$ is shown in Theorem 2, Φ and $\alpha/2$ are the cumulative density function and the upper $\alpha/2$th quantile of the standard normal distribution, respectively (Casella and Berger, 2001). We note that the power formula is a function of $\text{Var}(\hat{\theta})$. Following our discussion about the simplification of the variance formula (2.4) to the cases of arbitrary two-level CRTs in Hussey and Hughes (2007) and the three-level parallel CRT in Heo and Leon (2008), our power formula includes their works as special cases.

In practice, if the variance components are unknown, their ML estimators $\hat{\sigma}_a^2$, $\hat{\sigma}_b^2$ and $\hat{\sigma}_e^2$ can be substituted into the test statistic Z. In this case, we arrive at the Wald test statistic $W = \frac{\hat{\theta}-\theta_0}{\sqrt{\widehat{\text{Var}}(\hat{\theta})}}$, which follows the standard normal distribution asymptotically under H_0. The approximate power under θ_a can then be obtained by replacing $\text{Var}(\hat{\theta})$ with $\widehat{\text{Var}}(\hat{\theta})$ in (2.6).

In clinical research, it is usually more interesting to relate power to correlations among observations than to variance components. Thus, we introduce two concepts related to correlations defined under our model (2.1). The first quantity is called the intra-cluster correlation (ICC), $\rho \coloneqq \frac{\sigma_a^2+\sigma_b^2}{\sigma_a^2+\sigma_b^2+\sigma_e^2}$ (Heo and Leon, 2008). It is the correlation between two observations from the same unit within a cluster. The second quantity is $\eta \coloneqq \frac{\sigma_a^2}{\sigma_a^2+\sigma_b^2}$ (Hemming et al., 2015b), which is the ratio of the correlation between two observations from two different units within the same cluster to that from the same unit within a cluster. Next, let $\sigma_y^2 = \sigma_a^2 + \sigma_b^2 + \sigma_e^2$ denote the variance of the outcome Y_{ijtk}. To re-express the power formula in (2.6) in terms of ρ, η, σ_y^2 instead of $\sigma_a^2, \sigma_b^2, \sigma_e^2$, we rewrite $\text{Var}(\hat{\theta})$ in (2.4) in terms of the former three quantities

using the relation

$$
\begin{aligned}
p &= \frac{K}{(1-\rho)\sigma_y^2}, \\
q &= \frac{K(1-\eta)\rho}{1-\rho+TK(1-\eta)\rho}, \\
r &= \frac{K^2\eta\rho}{[1-\rho+TK(1-\eta)\rho+KJT\eta\rho][1-\rho+TK(1-\eta)\rho]\sigma_y^2}.
\end{aligned} \tag{2.7}
$$

Before a trial can be carried out, it is necessary to determine the number of participants to be recruited. Here, we briefly introduce the sample size calculations for arbitrary three-level CRTs based on the power formula (2.6). Given a clinically meaningful effect size $\theta_a - \theta_0$, the study planners would like to achieve a desired power π for rejecting the null hypothesis. By rearranging terms in (2.6), this criterion yields

$$\mathrm{Var}(\hat{\theta}) = (\frac{\theta_a - \theta_0}{z_{1-\pi} + z_{\alpha/2}})^2,$$

where $z_{1-\pi}$ and $z_{\alpha/2}$ are the upper $1-\pi$ and $\alpha/2$ quantiles of the standard normal distribution, respectively. Assuming that reliable estimates of ρ, η and σ_y^2 can be provided for a specific design Z_{ijt}, investigators may set the expression of $\mathrm{Var}(\hat{\theta})$ in (2.4) (with p, q, r expressed in (2.7)) equal to $(\frac{\theta_a-\theta_0}{z_{1-\pi}+z_{\alpha/2}})^2$, and solve for one of the four sizes I, J, T, K while having the other three prespecified. Following this procedure, we are able to determine the size of interest, I, J, T or K. In practice, common methods of estimating the ICC ρ and the ratio of between-unit to within-unit correlations η include referring to previous trials, conducting pilot studies, using baseline data, and carrying out interim analysis (Eldridge and Kerry, 2012).

2.3 Application

A stepped wedge study is planned to provide a newly designed training course to general practice nurses in type 2 diabetes care. Nurses who have received the new

training course are expected to provide better care and help reduce patients' total cholesterol level more than nurses who have not received the training. Following the balanced stepped wedge design shown in Figure 1.6 (d), the new training is delivered sequentially to general practices whose orders are randomly determined. The study consists of $T = 16$ phases, including an initial phase when nurses from all participating general practices deliver the regular care to their patients. After the initial phase of the study, it is planned that a fixed number of general practices transfer from the control arm to the intervention arm in each following phase. In addition, it is expected that $J = 3$ nurses are recruited from each general practice, and $K = 25$ patients can be recruited from each general practice at each phase. Patients' total cholesterol levels are collected two months after they receive either the regular care or the new care. Based on previous small studies, the estimated correlation ρ between patient observations from the same nurse is 0.05, and the estimated ratio η of between-nurse to within-nurse correlations is 0.3. Furthermore, the standard deviation of patients' total cholesterol is estimated to be 1.2 mmol/l. To detect a clinically important reduction 0.05 mmol/l in total cholesterol with 80% power, we can use the general power formula (2.6) with the above stepped wedge design plugged in. We have

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T} Z_{ijt}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T} Z_{ijt} = \frac{IJT}{2},$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(\sum_{t=1}^{T} Z_{ijt})^2 = \frac{IJT(2T-1)}{6},$$

$$\sum_{i=1}^{I}(\sum_{j=1}^{J}\sum_{t=1}^{T} Z_{ijt})^2 = \frac{IJ^2T(2T-1)}{6},$$

$$\sum_{t=1}^{T}(\sum_{i=1}^{I}\sum_{j=1}^{J} Z_{ijt})^2 = \frac{I^2J^2T(2T-1)}{6(T-1)}.$$

Substituting $J = 3, T = 16, K = 25, \rho = 0.05, \eta = 0.3$ and $\sigma_y = 1.2$ mmol/l into the variance formula (2.4) and power formula (2.6), we obtain the desired number of general practices I. It turns out that at least 45 general practices are needed in this SW-CRT. Specifically, during each of the 15 phases with switch of treatment arms, 3 general practices should be randomly chosen to transfer from the control to the intervention.

Suppose investigators would also like to know how many general practices would be needed if a two-arm parallel CRT were to be carried out, assuming all other conditions being same as above. In this case, we substitute the corresponding treatment indicator Z_{ijt} into the variance formula (2.4). Assuming the parallel design with a single time period ($T = 1$), we have

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T} Z_{ijt}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T} Z_{ijt} = \sum_{i=1}^{I}\sum_{j=1}^{J} Z_{ij} = \frac{IJ}{2},$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(\sum_{t=1}^{T} Z_{ijt})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} Z_{ij}^2 = \frac{IJ}{2},$$

$$\sum_{i=1}^{I}(\sum_{j=1}^{J}\sum_{t=1}^{T} Z_{ijt})^2 = \sum_{i=1}^{I}(\sum_{j=1}^{J} Z_{ij})^2 = \frac{IJ^2}{2},$$

$$\sum_{t=1}^{T}(\sum_{i=1}^{I}\sum_{j=1}^{J} Z_{ijt})^2 = (\sum_{i=1}^{I}\sum_{j=1}^{J} Z_{ij})^2 = (\frac{IJ}{2})^2.$$

The resulting formula of $\mathrm{Var}(\hat{\theta})$ is given in (2.5). Using the power formula (2.6), we decide the required number of clusters I plugging in the same values of $J, T, K, \rho, \eta, \sigma_y, \alpha$ and the clinically important reduction amount. It turns out that at least $I = 712$ general practices would be needed so that power achieves 80%. This number is substantially larger than the number of required general practices under a stepped wedge design. In fact, if only 46 general practice were to be recruited in a balanced parallel CRT, the power would only achieve 0.11.

2.4 Discussion

In clinical research, three-level data that include units nested within clusters are commonly observed. Examples include ICUs within hospitals, nurses within general practices, and villages within geographical areas. Model (2.1) including both cluster and random unit effects serves as a standard model to describe three-level CRTs. However, deriving the closed-form expression for the variance of the GLS estimator of the intervention effect and the corresponding power formula under model (2.1) is a nontrivial task because of the additional complexity in the covariance structure due to the random unit effect. The main contribution of this chapter – a closed-form expression for power of testing the intervention effect based on its GLS estimator under the standard three-level model (2.1) for arbitrary designs – is a significant leap forward on power and sample size calculations for three-level CRTs.

As a key component in obtaining our power formula, the derived closed-form expression for the variance of the GLS estimator $\hat{\theta}$ in (2.4) can lead to other analytic work. One important and interesting topic is to compute the relative efficiency of a design A to another design B, which is defined to be the ratio of the variance of the intervention effect estimator under design B to that under design A. Hence, our flexible variance formula will allow practitioners more flexibility in trial design and make easier to compare efficiencies of candidate designs.

Chapter 3: A Novel Set of Three-Level Stepped Wedge Cluster Randomized Trials

In this chapter, we propose a novel set of stepped wedge cluster randomized trials for three-level data such as patients (observations) within wards (units) within hospitals (clusters). The proposed SW-CRTs differ in timing of allocating different units from the same cluster to the intervention. Following our design concept, we provide four paradigms of three-level SW-CRTs, discuss their practical usage, and compare their efficiencies using the precision of the intervention effect estimator under the standard three-level model (2.1).

3.1 Stepped Wedge Designs for Three-Level Data

In traditional three-level SW-CRTs (as proposed by Hemming et al. (2015b)), clusters such as hospitals serve as the units of randomization. That is, all units (wards) from the same cluster (hospital) are supposed to complete the transition from the control to the intervention at the same time to avoid contamination between the units. However, when contamination is not a large concern, we can consider alternative ways of allocating the units within the same cluster. In the example of wards within hospitals, study planners could initially let only a few participating wards be exposed to the intervention. As the study proceeds, more and more wards make the

transition from the control to the intervention, and the time points when different wards within the same hospital switch to the intervention arm may be different.

Following the above idea of stepped wedge designs, we provide four paradigms of three-level SW-CRTs in Figure 3.1. Assume that we have 6 clusters (hospitals) and 6 units (wards) within each cluster. We focus on cross-sectional studies where a different set of patients are recruited in each unit at each time point.

- **Design 1 (All units transfer at a single time point)**. Proposed by Hemming et al. (2015b), this traditional stepped wedge design make all units from the same cluster complete the transition from the control to the intervention within a single time period. In Figure 3.1 (a), at each time point one of the 6 clusters is chosen to transfer to the intervention arm, leading to a study that lasts for 7 time periods and includes 6 steps of transitions. This type of design is most suitable when there is limitation on the number of practitioners and thus the intervention cannot be provided to too many clusters simultaneously.

- **Design 2 (Units transfer at two adjacent time points)**. In this design, units within the same cluster transfer from the control to the intervention at two separate and adjacent time points. In Figure 3.1 (b), 3 units (numbered 1–3) from each of clusters 1 and 2 complete the transition to the intervention at time point 2. Then the rest of the units (numbered 4–6) from the same two clusters complete the transition at time point 3. Following this procedure, units from clusters 3–6 have crossovers to the intervention arm at later time points so that eventually all participating clusters are exposed to the intervention. Compared to design 1, design 2 may require more practitioners, since the intervention is delivered to more clusters simultaneously.

- **Design 3 (Units transfer at two nonadjacent time points)**. Same as in design 2, each participating cluster completes the transition to the intervention in two separate time points. The difference is that the two time point when crossovers happen within the same cluster are not adjacent. In Figure 3.1 (c), 3 units (numbered 1–3) respectively selected from clusters 1 and 2 have crossovers to the intervention at time point 2. The rest of the units (numbered 4–6) from the two clusters are not exposed to the intervention until three periods after (time point 5). Same with clusters 1 and 2, clusters 3–6 complete the one-way crossover to the intervention in two nonadjacent time points. We note that the first three steps of the design among the 6 clusters is in fact a smaller-scale stepped wedge design. In this small-scale design, there are 6 clusters and 3 units within each cluster. Interim analysis could potentially be carried out in the middle of the study with sufficient number of independent clusters. Depending on whether the intervention turns out to be beneficial or harmful, investigators can decide to continue the study or terminate it earlier than planned.

- **Design 4 (Units transfer at all of the time points)**. In this design, study planners let all participating clusters have crossovers at all steps. Specifically, in Figure 3.1 (d), one unit from each cluster is selected to take the initial crossover to the intervention at time point 2. In time point 3, another unit from each of the cluster is selected and completes the transition. In this fashion, a cluster is not fully exposed to the intervention until the last time point. In this case, a mini stepped wedge design is planned for each participating cluster. Given adequate labor, a team can be divided into multiple groups, each of which work within a cluster until all units within the cluster are exposed to the intervention.

The advantage is that these small research groups do not need to travel among clusters, especially when traveling is difficult or costly.

The above four designs are fully balanced in the sense that each cluster contains the same number of units and the number of clusters making the crossover is the same at all steps. In other scenarios, such balance may not be maintained. For example, let us consider an unbalanced version of design 2, shown in Figure 3.2. At the initial stage of the design, three clusters (numbered 1–3) are selected from cluster 1 and make the transition to the intervention at time point 2. The rest of the units in cluster 1 (numbered 4–6), together with units 1–3 from cluster 2, are designed to transfer at time point 3. At time point 4, units 4–6 from cluster 2 and units 1–3 from cluster 3 are scheduled to make the transition to the intervention. Following this pattern, all participating clusters are exposed to the intervention at time point 8, where units 4–6 from the last cluster, cluster 6, finish the transition. This unbalanced version of design 2 is one period longer than the balanced version in Figure 3.1 due to the initial stage where only one cluster participates in the transition process. This unbalanced design provides more flexibility and ease to investigators, since it allows fewer participating clusters to transfer to the intervention at the initial stage of a study.

		Time						
Pattern	(Cluster, Units)	1	2	3	4	5	6	7
1	(1, 1-6)	0	1	1	1	1	1	1
2	(2, 1-6)	0	0	1	1	1	1	1
3	(3, 1-6)	0	0	0	1	1	1	1
4	(4, 1-6)	0	0	0	0	1	1	1
5	(5, 1-6)	0	0	0	0	0	1	1
6	(6, 1-6)	0	0	0	0	0	0	1

(a) Design 1: All units in the same cluster transfer at a single step

		Time						
Pattern	(Cluster, Units)	1	2	3	4	5	6	7
1	(1-2, 1-3)	0	1	1	1	1	1	1
2	(1-2, 4-6)	0	0	1	1	1	1	1
3	(3-4, 1-3)	0	0	0	1	1	1	1
4	(3-4, 4-6)	0	0	0	0	1	1	1
5	(5-6, 1-3)	0	0	0	0	0	1	1
6	(5-6, 4-6)	0	0	0	0	0	0	1

(b) Design 2: Units in the same cluster transfer within two adjacent steps

		Time						
Pattern	(Cluster, Units)	1	2	3	4	5	6	7
1	(1-2, 1-3)	0	1	1	1	1	1	1
2	(3-4, 1-3)	0	0	1	1	1	1	1
3	(5-6, 1-3)	0	0	0	1	1	1	1
4	(1-2, 4-6)	0	0	0	0	1	1	1
5	(3-4, 4-6)	0	0	0	0	0	1	1
6	(5-6, 4-6)	0	0	0	0	0	0	1

(c) Design 3: Units in the same cluster transfer in two nonadjacent steps

		Time						
Pattern	(Cluster, Units)	1	2	3	4	5	6	7
1	(1-6, 1)	0	1	1	1	1	1	1
2	(1-6, 2)	0	0	1	1	1	1	1
3	(1-6, 3)	0	0	0	1	1	1	1
4	(1-6, 4)	0	0	0	0	1	1	1
5	(1-6, 5)	0	0	0	0	0	1	1
6	(1-6, 6)	0	0	0	0	0	0	1

(d) Design 4: Units in the same cluster transfer at all of the steps

Figure 3.1: Examples of complete SW-CRTs. In each cell, 0 means receiving the control, and 1 means receiving the intervention. Clusters are numbered from 1 to 6. Units within each cluster are also numbered 1–6. The (cluster, unit) pair indicates which unit from which cluster. For example, (2, 1-3) means units 1–3 from the second cluster, and (1-6, 1) means the first unit from each of the six clusters.

		Time							
Pattern	(Cluster, Units)	1	2	3	4	5	6	7	8
1	(1, 1-3)	0	1	1	1	1	1	1	1
2	(1, 4-6); (2, 1-3)	0	0	1	1	1	1	1	1
3	(2, 4-6); (3, 1-3)	0	0	0	1	1	1	1	1
4	(3, 4-6); (4, 1-3)	0	0	0	0	1	1	1	1
5	(4, 4-6); (5, 1-3)	0	0	0	0	0	1	1	1
6	(5, 4-6); (6, 1-3)	0	0	0	0	0	0	1	1
7	(6, 4-6)	0	0	0	0	0	0	0	1

Figure 3.2: An unbalanced version of design 2. In each cell, 0 means receiving the control, and 1 means receiving the intervention.

3.2 Efficiencies of Proposed Designs under the Standard Model

We compare the variances of the intervention effect estimator $\hat{\theta}$ for designs 1 to 4 in Figure 3.1 under the standard three-level model (2.1) to investigate the efficiencies of these designs. Recall that model (2.1) is

$$Y_{ijtk} = \delta_k + Z_{ijt}\theta + a_i + b_{ij} + \epsilon_{ijtk},$$

where $i = 1, \cdots, I$ is the index for independent clusters, $j = 1, \cdots, J$ is the index for units nested within the same cluster, $t = 1, \cdots, T$ is the index for time points, and $K = 1, \cdots, K$ is the index for observations from the same unit within a cluster at the same time. Z_{ijt} is the binary intervention indicator for the jth unit within the ith cluster at time t, taking the value of 1 if the intervention is assigned, and 0 otherwise. δ_t and θ denote the fixed time effect at time t and the intervention effect, respectively. The cluster random effects $a_i, i = 1, \cdots, I$ are independently drawn from $N(0, \sigma_a^2)$. Within the same cluster i, the random unit effects $b_{ij}, j = 1, \cdots, J$ are independently drawn from $N(0, \sigma_b^2)$. The random errors ϵ_{ijtk} are independently drawn from $N(0, \sigma_e^2)$.

Based on model (2.1), recall that the two concepts related to the correlations are defined. First, the intra-cluster correlation (ICC) is defined to be $\rho \coloneqq \frac{\sigma_a^2+\sigma_b^2}{\sigma_a^2+\sigma_b^2+\sigma_\epsilon^2}$ (Heo and Leon, 2008). It is simply the correlation between two observations from the same unit within a cluster. Second, the ratio of the between-cluster to within-cluster correlations is $\eta \coloneqq \frac{\sigma_a^2}{\sigma_a^2+\sigma_b^2}$ (Hemming et al., 2015b). It represents the relative contribution of the clusters to the correlation between two observations within the same unit in a cluster.

3.2.1 Study Setup

Assume that there are $I = 18$ participating clusters, with $J = 6$ units in each cluster. For each of the four designs, the length of the study is $T = 7$ time periods. We follow a fully balanced design. Hence, under design 1, there are 3 clusters included in each transition step. Under designs 2 and 3, there are 6 clusters included in each transition step. Under design 4, all 18 clusters are included in each transition step. The number of observations taken from each unit within a cluster at each time point is $K = 20$. The variance of Y_{ijtk} is assumed to be 5.

Using the formula of $\text{Var}(\hat{\theta})$ in Theorem 2, we compare the variances of the intervention effect estimator $\hat{\theta}$ across different designs and study the pattern in $\text{Var}(\hat{\theta})$ as ρ or η varies. When η increases from 0 to 1, ρ is fixed at 0.001, 0.01, 0.1, and 0.2, respectively. On the other hand, when ρ increases from 0 to 0.6, η is fixed at 0, 0.3, 0.7, and 1, respectively.

3.2.2 Different Types of Comparisons

Before introducing the results of the study, we would like to pause and introduce three different possible comparisons between the intervention arm and the control arm under model (2.1):

1. Within-unit Between-time comparisons. These are comparisons of observations from the same unit within the same cluster between different time points. We expect this type of comparisons to be most efficient when the correlation ρ of observations within the same unit is large. This situation is analogous to crossover trials, where each cluster serves as its own control.

2. Between-unit within-cluster comparisons. These are comparisons of observations from different units within the same cluster.We expect this type of comparisons to be most efficient when η is large for fixed ρ and σ_y^2. In that case, units from the same cluster are more similar and thus more able to serve as controls for each other. This type of comparison may include both comparisons at the same time point and comparisons at two different time points. Note that the comparisons at the same time point only contribute to the estimation of θ in designs 2, 3 and 4.

3. Between-cluster comparisons. These are comparisons of observations from different clusters receiving different treatments. We expect this type of comparisons to be most efficient when both ρ and η are small. In that case, observations from the same cluster are close to being independent, and this type of comparison is close to the comparison made between two treatment arms in an

Table 3.1: Change in efficiencies of different comparisons for different values of ρ and η under model (2.1)

Conditions \ Comparisons	Within-unit Between-time	Between-unit Within-cluster	Between-cluster
η constant, ρ increases	increase	increase	decrease
ρ constant, η increases	same	increase	decrease

individually randomized trial. In addition, this type of comparison may include both comparisons at the same time and those at two different time points.

Table 3.1 summarizes different contributions of the three types of comparisons to the precision of the treatment effect estimator $\hat{\theta}$ under the standard three-level model (2.1) when either ρ or η varies.

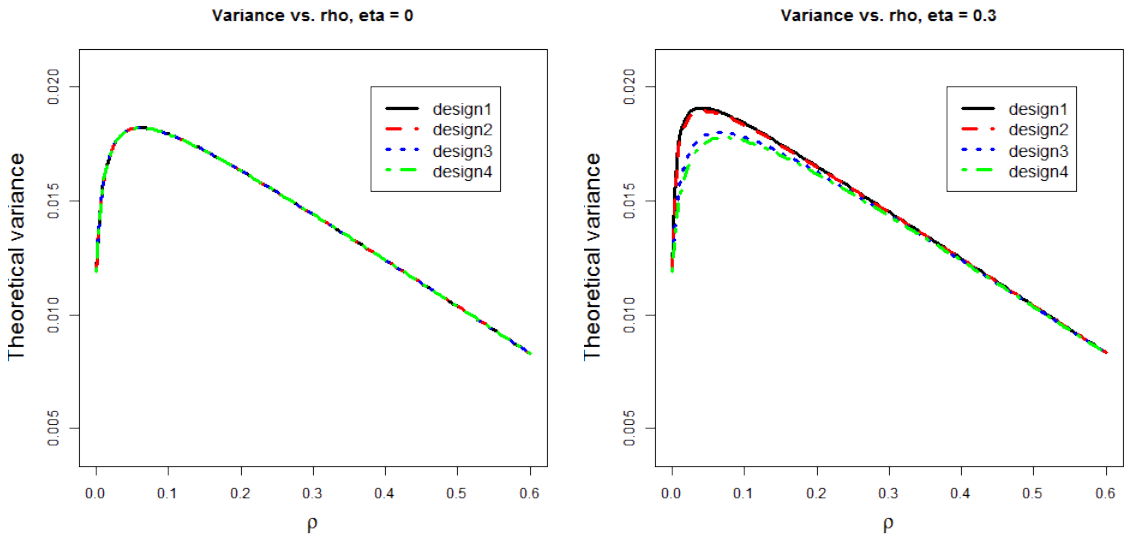
3.2.3 Results

When η is held fixed and ρ increases from 0 to a small threshold, one tends to lose precision of $\hat{\theta}$ due to worsening effects of between-cluster comparisons (type 3) that outweigh positive effects of within-cluster comparisons (types 1 and 2) (Figure 3.3). When ρ passes through some threshold, the effects of the within-cluster comparisons become stronger than the between-cluster comparisons. Thus, the precision of $\hat{\theta}$ is gradually improved, and the variance of $\hat{\theta}$ ($\text{Var}(\hat{\theta})$) decreases. Hence, in each panel of Figure 3.3, the theoretical variance $\text{Var}(\hat{\theta})$ first increases as ρ increases within some threshold, and then decreases after ρ passing the threshold.

It is also worth noticing that when ρ is fixed and constant, the difference in $\text{Var}(\hat{\theta})$ among the four designs becomes larger and larger as η increases (Figure 3.3). This

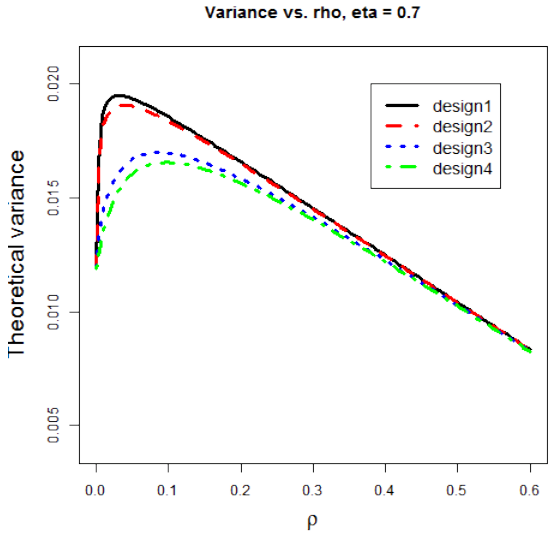
difference can be explained by the between-unit within-cluster within-time comparisons. Since designs 2, 3, and 4 contain this type of comparisons in estimating θ, they accumulate precision faster than design 1 as η increases. Furthermore, design 4 gains the highest precision due to having the most between-unit within-cluster comparisons at the same time point. We also note that when $\eta = 0$, the four designs are equivalent, since all participating units are now independent and serve as independent "clusters". Thus, the four lines completely overlap with one another in Figure 3.3 (a).

As another way to consider the impact of η, we provide plots with ρ fixed and η increasing, shown in Figure 3.4. In this scenario, the effect of within-unit between-time comparisons (comparison 1) is constant, while the effects of the other two comparisons depend on the magnitude of η. As η increases, the precision of $\hat{\theta}$ can either increase or decrease. With increasing η, the precision of the between-unit within-cluster comparisons (type 2) improve precision and the precision of the between-cluster comparisons (type 3) decrease. Thus, because of the difference in the number of between-unit within-cluster comparisons across designs, the theoretical variance show quite different patterns under designs 1–4 with increasing η.

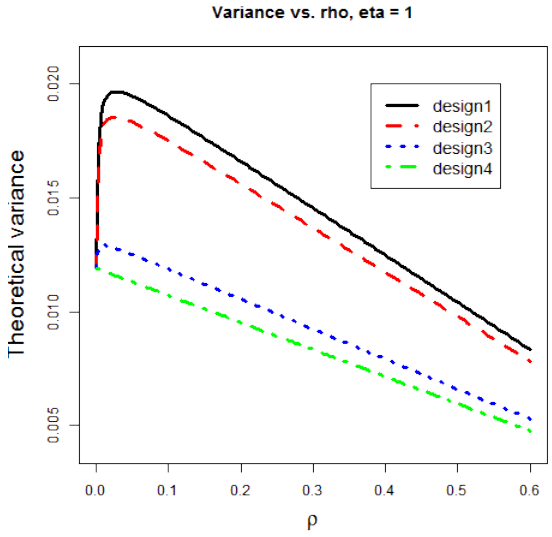


(a) $\eta = 0$

(b) $\eta = 0.3$

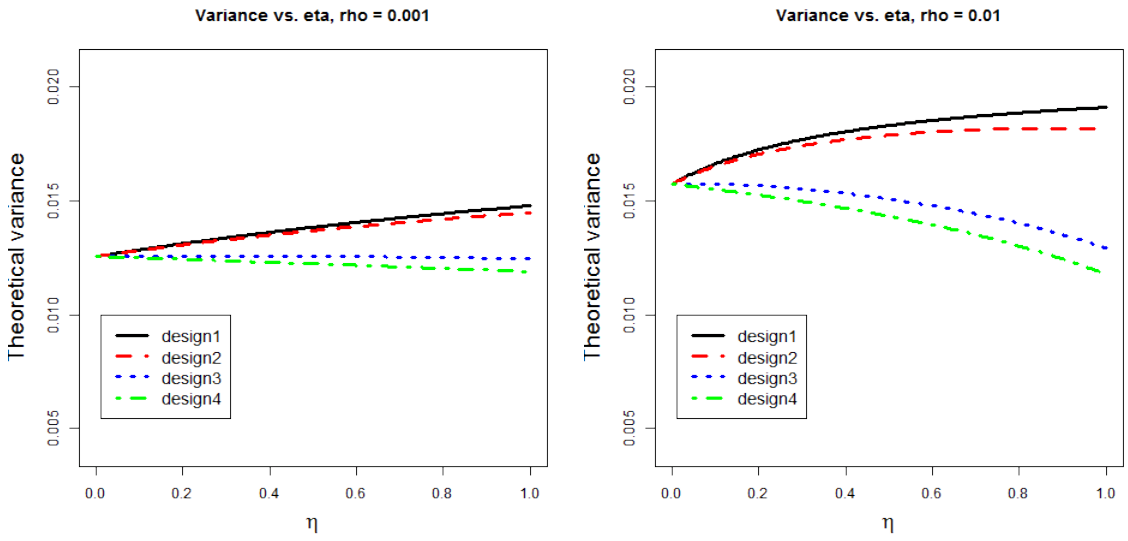


(c) $\eta = 0.7$



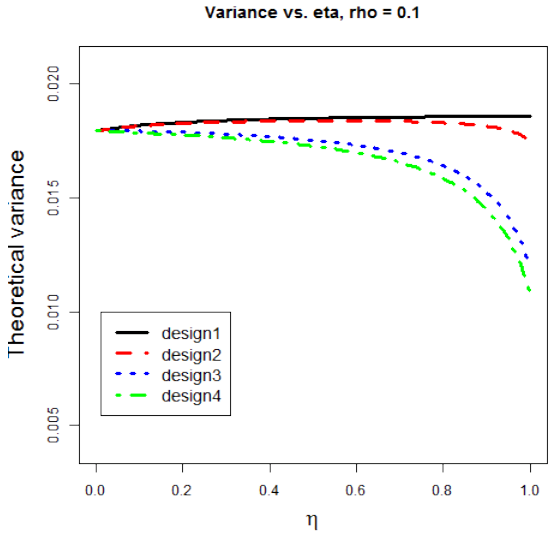
(d) $\eta = 1$

Figure 3.3: $\mathrm{Var}(\hat{\theta})$ vs. ρ. In each of the four panels, η is fixed at 0, 0.4, 0.7, and 1, respectively.

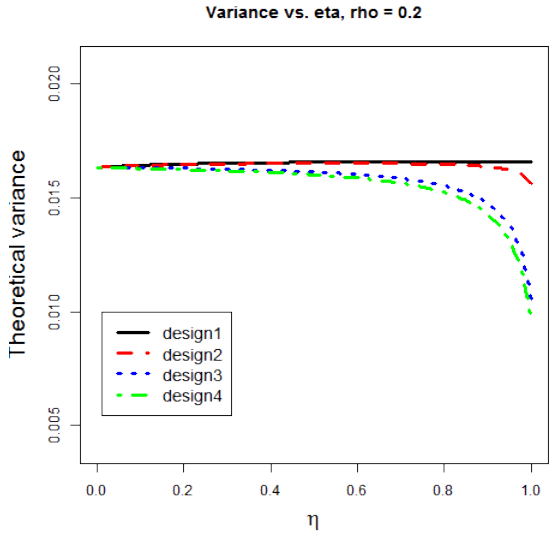


(a) $\rho = 0.001$

(b) $\rho = 0.01$



(c) $\rho = 0.1$



(d) $\rho = 0.2$

Figure 3.4: $\mathrm{Var}(\hat{\theta})$ vs. η. In each of the four panels, ρ is fixed at 0.001, 0.01, 0.1, and 0.2, respectively.

There is an issue with residual degrees of freedom when the number of clusters I is small and the correlation ρ between individuals from the same unit in a cluster is big. While large values of ρ and η lower the precision of between-cluster comparisons (type 3) and thus may require larger sample size, the one-way crossover within each unit/cluster improves the precision via within-cluster comparisons (types 1 and 2). Hence, the degrees of freedom should not be a large concern under the proposed three-level SW-CRTs. We conducted a small simulation study to compute the denominator degrees of freedom in the t test for $H_0 : \theta = 0$ for all designs. We set $\rho = 0.99$, $\eta = 0.3$, $I = 6$, $J = 6$, $T = 7$, and $K = 5$. We also assumed $\theta = 0.4, \sigma_y = 5$ and no time effects, so $\delta_1 = \cdots = \delta_T = 50$. We generated 100 datasets using the standard three-level model (2.1) under each design. When fitting the data using the same model, we employed Kenward-Roger degrees of freedom approximation. For designs 1–4, we obtained the denominator degrees of freedom $1217, 1206, 1204$ and 1204, respectively. Hence, in the cases of three-level SW-CRTs, denominator degrees of freedom may not be a large concern even when correlations among observations within the same cluster or unit are large.

3.3 Discussion

In this chapter, we proposed a set of novel SW-CRTs for three-level data such as patients within wards within hospitals. The designs mainly differ in timing of assigning different units within the same cluster to the intervention. We provided schemes of three-level SW-CRTs to illustrate this idea and discussed the practical implications of the proposed designs. By comparing the theoretical variances under the four designs given various combinations of ρ and η, we conclude that design 1 is

the least efficient, and design 4 is the most efficient. Thus, a stepped wedge design that spreads units within the same cluster over steps would yield the highest precision in the estimation of the intervention effect, assuming no contamination.

The four proposed designs may be employed under different scenarios. Design 1 has simplest form among the four designs. Thus, it can be more easily understood and implemented. It requires fewer investigators but potentially higher traveling cost, since a research team needs to travel from cluster to cluster to roll out the intervention. Design 2 is useful when study planners prefer that the rest of the units within a cluster need to crossover to the intervention soon after the first few units take the crossover. In community-based CRTs, this design may be chosen for political reasons. It would be easier to encourage communities to participate if the intervention is promised to deliver sooner within one community. Design 2 requires an adequate number of practitioners and traveling expenses, since there should be multiple research teams at each transition step. In design 3, each cluster completes the crossover within two non-adjacent steps. It yields a small-scale SW-CRT in the first half of the design. This design would allow interim analysis. In addition, design 3 require a moderate amount of labor and traveling expenses. Finally, design 4 requires the highest amount of labor but the least amount of traveling. This is because the number of research teams is required to be equal to the number of clusters, but practitioners do not need to travel among clusters.

Chapter 4: Model Misspecification under Three-Level SW-CRTs

In Chapters 2 and 3, we modeled three-level CRTs using the standard three-level model (2.1) where the random unit effect b_{ij} is constant across all time points, as might be true if the unit effects are related to factors that do not change with time. However, this may not always be true for real-world data collected from SW-CRTs that typically last for long periods of time. In this chapter, we consider an underlying three-level model to generate data with random unit effects varying across time points. In the example of nurses within general practices provided in Section 2.3, the effect of a nurse on an outcome variable at one time point may not be the same as the effect of the same nurse at a different time point. For example, patient-nurse interaction may change as time evolves, which in turn may affect patients responses over time. For the jth nurse from the ith general practice, the nurse random effect at time t, b_{ijt}, may not be equal to that at time t', $b_{ijt'}$, $t \neq t'$. The underlying model to generate data incorporates correlated random unit effects $b_{ijt}, t = 1, \cdots, T$.

This chapter is organized as follows. In Section 4.1, we introduce the underlying model with covariance pattern of b_{ijt} specified as Toeplitz. In Section 4.2, we consider the autoregressive covariance pattern as a special case of the Toeplitz covariance pattern. Given that the standard model (2.1) is a popular modeling choice,

we evaluate the impact of misspecifying the random unit effects $b_{ijt}, t = 1, \cdots, T$ to be constant across time. Also, to better understand this more complex underlying model, we study the precision of the treatment effect estimator when the model is correctly specified under each design. In Section 4.3, we consider general Toeplitz covariance patterns. As with Section 4.2, this section studies the precision of the treatment effect estimator when the data-generating model is correctly specified and examines the consequence of misspecifying random unit effects to be constant across time under designs 1–4. Section 4.4 includes discussion of this chapter.

4.1 The Model

Let us consider the following mixed effects model

$$Y_{ijtk} = \delta_t + Z_{ijt}\theta + a_i + b_{ijt} + \epsilon_{ijtk}, \tag{4.1}$$

where the notations follow from the standard three-level model (2.1). The index $i = 1, \cdots, I$ is for independent clusters, index $j = 1, \cdots, J$ is for units nested within the same cluster, index $t = 1, \cdots, T$ is for time points, and index $k = 1, \cdots, K$ is for observations from the same unit within a cluster at the same time. Accordingly, Y_{ijtk} denotes the kth observation from the jth unit nested in the ith cluster at time t. Z_{ijt} is the binary intervention indicator for the jth unit within the ith cluster at time t. It takes value of 1 if the intervention is received, and 0 if the control is received. δ_t and θ denote the fixed time effect at time t and the intervention effect, respectively. The random cluster effects $a_i, i = 1, \cdots, I$ are independently drawn from $N(0, \sigma_a^2)$, and the random errors ϵ_{ijtk} are independently drawn from $N(0, \sigma_e^2)$. The difference between model (4.1) and the standard model (2.1) is the random unit effects. In the standard model, the random unit effects $b_{ij}, j = 1, \cdots, J$ are

independently drawn from $N(0, \sigma_b^2)$. Here in model (4.1), we define the random vector $b_{ij} \coloneqq (b_{ij1}, \cdots, b_{ijT})^T \in \mathbb{R}^T$ to be the random effects of unit j from cluster i. They are independently and identically drawn from a $N_T(0, V_b)$ distribution, where V_b is a symmetric positive definite matrix. Note that a_i, b_{ij}, and ϵ_{ijtk} are mutually independent. As with previous sections, we focus on cross sectional studies where the observations are taken on $IJTK$ different participants.

We model the covariance of b_{ij} using the Toeplitz matrix which assumes constant variance σ_b^2 across different times, and $\text{Corr}(b_{ijt}, b_{ijt'}) = \phi_{b,|t-t'|}$ for any $t \neq t', t, t' = 1, \cdots, T$ and $\phi_{b,l} \in [0, 1]$. That is,

$$V_b = V_b(\sigma_b^2, \phi_{b,1}, \cdots, \phi_{b,T-1}) = \sigma_b^2 \begin{pmatrix} 1 & \phi_{b,1} & \phi_{b,2} & \cdots & \phi_{b,T-1} \\ \phi_{b,1} & 1 & \phi_{b,1} & \cdots & \phi_{b,T-2} \\ \phi_{b,2} & \phi_{b,1} & 1 & \cdots & \phi_{b,T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{b,T-1} & \phi_{b,T-2} & \phi_{b,T-3} & \cdots & 1 \end{pmatrix}. \quad (4.2)$$

Since the correlation between any two adjacent terms is assumed to be $\phi_{b,1}$, it requires that the time points at which observations are collected be equally spaced (Fitzmaurice et al., 2011), which is a reasonable assumption under the stepped wedge design.

Under some special conditions of $\phi_{b,1}, \cdots, \phi_{b,T-1}$, model (4.1) simplifies to some other models. First, when $\phi_{b,1} = \cdots = \phi_{b,T-1} = 1$, model (4.1) is essentially equivalent to the standard three-level model (2.1). On the other hand, when $\phi_{b,1} = \cdots = \phi_{b,T-1} = 0$, observations from the same unit within a cluster at different time points are correlated as closely as those from different units within the same cluster; there is only increased similarity of observations from the same unit taken at the same time.

Let us take a close look at the covariance matrix V of the vector of outcome $y \coloneqq (Y_{1111}, \cdots, Y_{111K}, Y_{1121}, \cdots, Y_{11TK}, Y_{1211}, \cdots, Y_{1JTK}, Y_{2111}, \cdots, Y_{IJTK})^T \in \mathbb{R}^{IJTK}$ under model (4.1). According to the distributions of the random effects a_i, b_{ijt} and the random error e_{ijtk}, we have the following:

- Variance of the outcome $\text{Var}(Y_{ijtk}) = \sigma_a^2 + \sigma_b^2 + \sigma_e^2$.
- Covariance between outcomes within the same cluster, same unit and same time $\text{Cov}(Y_{ijtk}, Y_{ijtk'}) = \sigma_a^2 + \sigma_b^2, k \neq k'$.
- Covariance between outcomes from the same cluster, same unit, and different times $\text{Cov}(Y_{ijtk}, Y_{ijt'k'}) = \sigma_a^2 + \sigma_b^2 \phi_{b,|t-t'|}, t \neq t', k \neq k'$.
- Covariance between outcomes from the same cluster and different units $\text{Cov}(Y_{ijtk}, Y_{ij'tk'}) = \text{Cov}(Y_{ijtk}, Y_{ij't'k'}) = \sigma_a^2, j \neq j', t \neq t', k \neq k'$.
- Covariance of outcomes between different clusters $\text{Cov}(Y_{ijtk}, Y_{i'j'tk'}) = \text{Cov}(Y_{ijtk}, Y_{i'j't'k'}) = 0, i \neq i', j \neq j', t \neq t', k \neq k'$.

Thus, the covariance matrix of the vector of outcome y is $V = I_I \otimes W$, where $W \in \mathbb{R}^{JTN \times JTN}$ denotes the covariance matrix of the outcome Y_{ijtk} for any fixed cluster i. For easy explanation, Figure 4.1 provides a visualization of W. In the figure, we assumed that $J = 3$ units are nested within each cluster, $T = 4$ time periods are covered by the study, and $K = 5$ observations are collected from each unit during each time period. The black diagonal entries denote the value of $\text{Var}(Y_{ijtk})$ that are largest among all the entries, the darkest gray diagonal blocks of size 5×5 denote entries of $\text{Cov}(Y_{ijtk}, Y_{ijtk'}), k \neq k'$, the lighter gray blocks of size 5×5 denote entries of $\text{Cov}(Y_{ijtk}, Y_{ijt'k'}), t \neq t', k \neq k'$, and the lightest gray entries in the off-diagonal blocks

indicate $\text{Cov}(Y_{ijtk}, Y_{ij'tk'})$ and $\text{Cov}(Y_{ijtk}, Y_{ij't'k'}), j \neq j', k \neq k'$, which are smallest among all the entries in W.

Based on model (4.1), we retain the two important concepts introduced under the standard three-level model (2.1) with minimal modification. First, the ICC among observations within the same unit of a cluster at the same time point is $\rho = \frac{\sigma_a^2+\sigma_b^2}{\sigma_a^2+\sigma_b^2+\sigma_e^2}$. Second, the ratio of the between-unit to within-unit within-time correlations is $\eta = \frac{\sigma_a^2}{\sigma_a^2+\sigma_b^2}$.

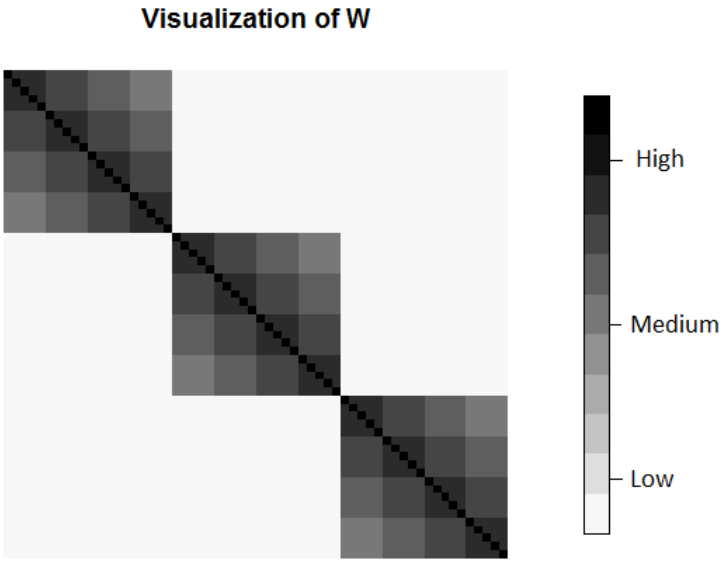


Figure 4.1: Visualization of covariance matrix W

4.2 Correlated random unit effects with the Autoregressive Covariance Pattern

As a special case of the Toeplitz covariance pattern, the autoregressive model for the covariance matrix assumes constant variance σ_b^2 across different times points

and $\text{Corr}(b_{ijt}, b_{ijt'}) = \phi_b^{|t-t'|}$ for $t, t' = 1, \cdots, T, t \neq t'$ and $\phi_b \in [0, 1]$. Compared to the general Toeplitz covariance pattern, the autoregressive covariance pattern is more parsimonious and only contains 2 parameters regardless of the number of time points.

The explicit form of the autoregressive model for the covariance of $b_{ij}, t = 1, \cdots, T$ is

$$V_b = V_b(\sigma_b^2, \phi_b) = \sigma_b^2 \begin{pmatrix} 1 & \phi_b & \phi_b^2 & \cdots & \phi_b^{T-1} \\ \phi_b & 1 & \phi_b & \cdots & \phi_b^{T-2} \\ \phi_b^2 & \phi_b & 1 & \cdots & \phi_b^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_b^{T-1} & \phi_b^{T-2} & \phi_b^{T-3} & \cdots & 1 \end{pmatrix}. \tag{4.3}$$

Hence, for two observations from the same unit within a cluster and different times, the covariance is $\text{Cov}(Y_{ijtk}, Y_{ijt'k'}) = \sigma_a^2 + \sigma_b^2 \phi_b^{|t-t'|}$ for any $t \neq t', t, t' = 1, \cdots, T$.

4.2.1 Variance of the Intervention Effect Estimator under the Correct Model

In this section, we study the patterns of the variance of the GLS estimator of the intervention effect θ with different combinations of ρ, η and ϕ_b when model (4.1) is correctly specified. The GLS estimator of θ is obtained by computing $(X^T V^{-1} X)^{-1}$, where X and V are the design matrix and the covariance matrix of the outcome under model (4.1), respectively.

Study Setup

As before, we let the number of clusters $I = 18$, the number of units within each cluster $J = 6$, the number of time points $T = 7$, and the cell size $K = 20$. The size of the intervention effect is assumed to be $\theta = 0.3$ so that the power of detecting such effect size achieves 80% when $\rho = 0$ under design 1. To simplify interpretation of results, we did not include any time effect in this study and let $\delta_1 = \ldots = \delta_7 = 50$ in model (4.1). In addition, we let the total variation of the outcome Y_{ijtk} be $\sigma_y = 5$.

The GLS estimator $\hat{\theta}$ under the correct model (4.1) is computed numerically using R software package (R Core Team, 2016).

The three quantities ρ, η and ϕ_b vary from 0 to 1 while the other two are held constant. We considered three scenarios as follows.

- When ρ increases, η varies among $0.01, 0.3, 0.6, 0.9$ for each fixed ϕ_b; ϕ_b varies among $0.01, 0.2, 0.5, 0.8, 0.99$ for each fixed η.
- When η increases, ρ varies among $0.01, 0.1, 0.2, 0.4$ for each fixed ϕ_b; ϕ_b varies among $0.01, 0.2, 0.5, 0.8, 0.99$ for each fixed ρ.
- When ϕ_b increases, ρ varies among $0.01, 0.1, 0.2, 0.4$ for each fixed η; η varies among $0.01, 0.3, 0.6, 0.9$ for each fixed ρ.

Different types of comparisons

As with the standard three-level model (2.1) in Chapter 3, the treatment effect is composed of three possible comparisons under model (4.1).

1. Within-unit between-time comparisons. These are comparisons of observations from the same unit within the same cluster between different times. We expect this type of comparison to be most beneficial when both the correlation ρ of observations within the same unit and the correlation ϕ_b between two adjacent random unit effects are large. Furthermore, when ρ is fixed, ϕ_b serves as a tuning parameter for how closely observations from the same unit between different times are correlated. Large values of ϕ_b yield large within-unit between-time correlations. In this case, each cluster serves as its own control, which is analogous to crossover trials.

2. Between-unit within-cluster comparisons. These are comparisons of observations from different units within the same cluster. We expect this type of comparisons to be most efficient when η is large for fixed ρ and σ_y^2. In that case, units from the same cluster are more similar and thus more able to serve as controls for each other. This type of comparison may include both comparisons at the same time point and comparisons at two different time points. Note that the comparisons at the same time point only contribute to the estimation of θ in designs 2, 3 and 4. In addition, the impact of ϕ_b on precision is minimal via these between-unit within-cluster comparisons.

3. Between-cluster comparisons. These are comparisons of observations from different clusters receiving different treatments. We expect this type of comparisons to be most efficient when ρ, η and ϕ_b are all small. In that case, observations from the same cluster are close to being independent, and this type of comparison is close to the comparison made between two treatment arms in an individually randomized trial. In addition, this type of comparison may include both comparisons at the same time and those at two different time points.

Table 4.1 summarizes different contributions of the three types of comparisons to the precision of the treatment effect estimator $\hat{\theta}$ under model (4.1) when one of ρ, η and ϕ_b varies and the other two are fixed.

Results

Our results contain three parts based on the scenarios considered. First, Figures 4.2 and 4.3 display the variance of the ML estimator $\hat{\theta}$ under model (4.1) for varying ρ and fixed η, ϕ_b. Second, Figures 4.4 and 4.5 display the variance of $\hat{\theta}$ under model

Table 4.1: Change in efficiencies of different comparisons for different values of ρ, η and ϕ_b under model (4.1)

Conditions \ Comparisons	Within-unit Between-time	Between-unit Within-cluster	Between-cluster
η, ϕ_b constant, ρ increases	increase	increase	decrease
ρ, ϕ_b constant, η increases	same	increase	decrease
ρ, η constant, ϕ_b increases	increase	same	decrease

(4.1) for varying η and fixed ρ, ϕ_b. Third, Figures 4.6 and 4.7 display the variance of $\hat{\theta}$ under model (4.1) for varying ϕ_b and fixed ρ, η. For ease of illustration, we hereby only show results under more extreme conditions. Results under less extreme conditions are in Appendix B.

We begin with the relation between $\text{Var}(\hat{\theta})$ and ρ. When ρ increases for fixed η and ϕ_b, one can either lose or gain precision of $\hat{\theta}$ depending on the magnitudes of η and ϕ_b (Figures 4.2 and 4.3). When ϕ_b is fixed at 0.99, as mentioned before, model (4.1) is nearly equivalent to the standard three-level model (2.1), since the random unit effects $b_{ijt}, t = 1, \cdots, T$ are highly correlated. We can follow similar reasoning for what we observed in Figure 3.3 in the previous chapter. As ρ increases from 0 to some small threshold, the variance of $\hat{\theta}$ increases due to worsening effects of between-cluster comparisons (type 3). Then, as ρ passes through the threshold, one tends to gain precision due to stronger effects of within-cluster comparisons (types 1 and 2), and thus the variance of $\hat{\theta}$ decreases. On the other hand, when ϕ_b is extremely small (0.01), the effect of within-unit between-time comparisons (type 1) depends primarily on the magnitude of η instead of ρ, and the effects of between-cluster comparisons

(type 3) keeps lowering the precision as ρ increases. Hence, in all the four panels in Figure 4.2, the variances are monotonically increasing as ρ increases.

Now let us investigate the relation between $\mathrm{Var}(\hat{\theta})$ and η. As with the case when ρ varies, one can either lose or gain precision of $\hat{\theta}$ given varying η, depending on the magnitudes of ϕ_b and ρ. Figure 4.4 provides the scenario when ϕ_b is extremely small (0.01), and ρ is fixed at $0.01, 0.1, 0.2, 0.4$, respectively. When both ϕ_b and ρ are small (panel (a)), observations from the same cluster are close to being independent and η does not have much effect on precision. Hence, we observe almost flat curves. As ρ is fixed at larger values (panels (b)–(d)), situations are different. Since the tuning parameter ϕ_b for the within-cluster between-time correlation is small, the within-unit between-time comparisons (type 1) would be most beneficial given large values of η for a fixed ICC ρ. When η increases from 0 to some small threshold, one tends to lose precision under designs 1 and 2 due to worsening effects of between-cluster comparisons (type 3). As η passes through some threshold, the variance begins to decrease and thus the precision increases due to stronger effect of within-cluster comparisons (types 1 and 2) under designs 1 and 2. Since designs 3 and 4 have more between-unit within-time comparisons (included in type 2 comparisons), they never lose (and keeps accumulating) precision as η increases.

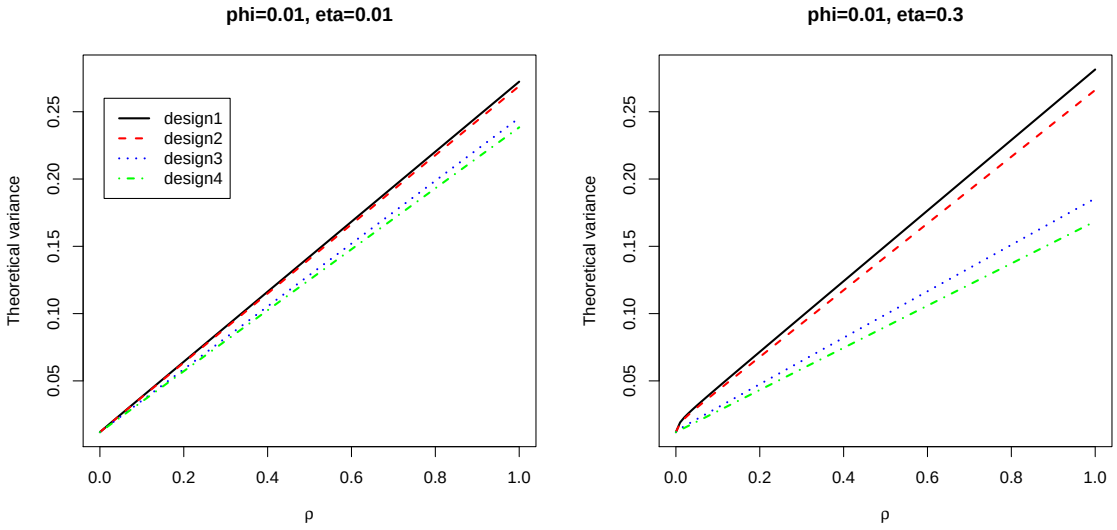
Providing another extreme scenario, Figure 4.5 shows results when ϕ_b is large (0.99). In this case, model (4.1) is nearly equivalent to the standard three-level model (2.1). Thus, similar reasoning as in Section 3.2 can be applied here. Given large values of η, precision increases through between-unit within-cluster comparisons (type 2). In addition, as η increases, designs 3 and 4 accumulate precision much faster

than designs 1 and 2, since they contain more between-unit within-time comparisons (included in type 2 comparisons).

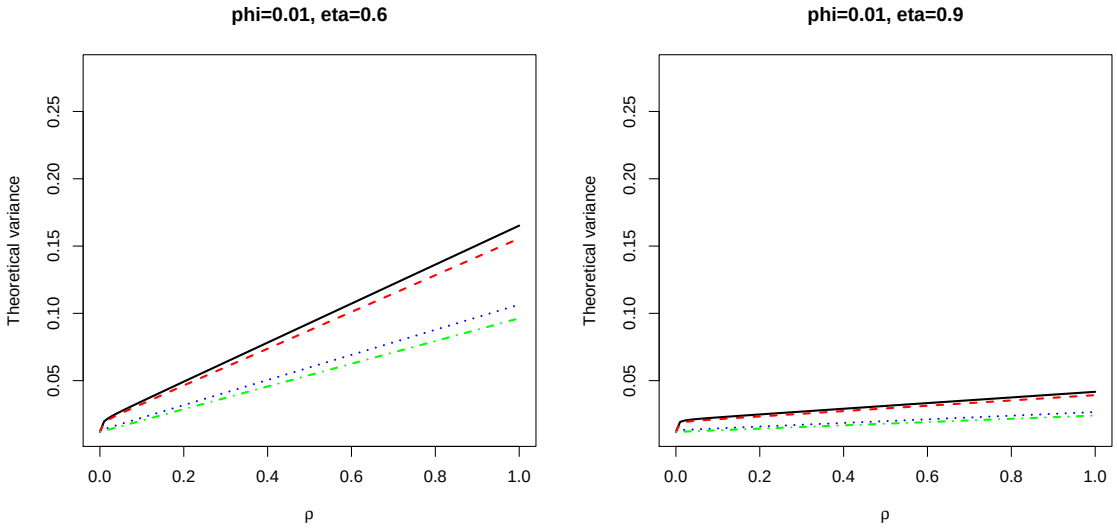
The last set of scenarios that we consider is to vary ϕ_b while keeping ρ and η fixed. When both ρ and η are small (Figure 4.6 (a)), observations within the same cluster are close to independent. In that case, varying ϕ_b does not yield large differences in estimation precision. As the ICC ρ is fixed at larger values, an increasing value of ϕ_b together with the large ICC ρ would make the within-unit between-time comparisons (type 1) more and more beneficial. Within a threshold of ϕ_b, the worsening effects of between-cluster comparisons (type 3) are dominating as ϕ_b increases. Beyond the threshold, the effect of within-unit between-time comparisons (type 1) gradually improves and is instead dominating. Thus, we observe upside-down U-shaped patterns in the variance of $\hat{\theta}$ in Figure 4.6 (b)–(d). Note that in this scenario, the effects of between-unit within-cluster comparisons (type 2) are small due to the small value of η in this case.

Lastly, we consider a large value of η while varying ϕ_b. When the ICC ρ is extremely small, even a large value of ϕ_b may not make the within-unit between-time comparisons (type 1) very beneficial. Thus, we observe nearly flat curves in Figure 4.7 (a), where designs 3 and 4 are still more efficient due to more between-unit within-time comparisons (included in type 2 comparisons) given the large magnitude of η. In the other three panels in Figure 4.7, the worsening effects of between-cluster comparisons (type 3) are dominating within a threshold of ϕ_b and lead to increase in $\text{Var}(\hat{\theta})$, and the beneficial effects of within-cluster comparisons (types 1 and 2) are dominating beyond the threshold and lead to decrease in $\text{Var}(\hat{\theta})$. With a large value

of η, the precision increases very fast due to beneficial impact of the between-unit within-cluster comparisons (type 2).

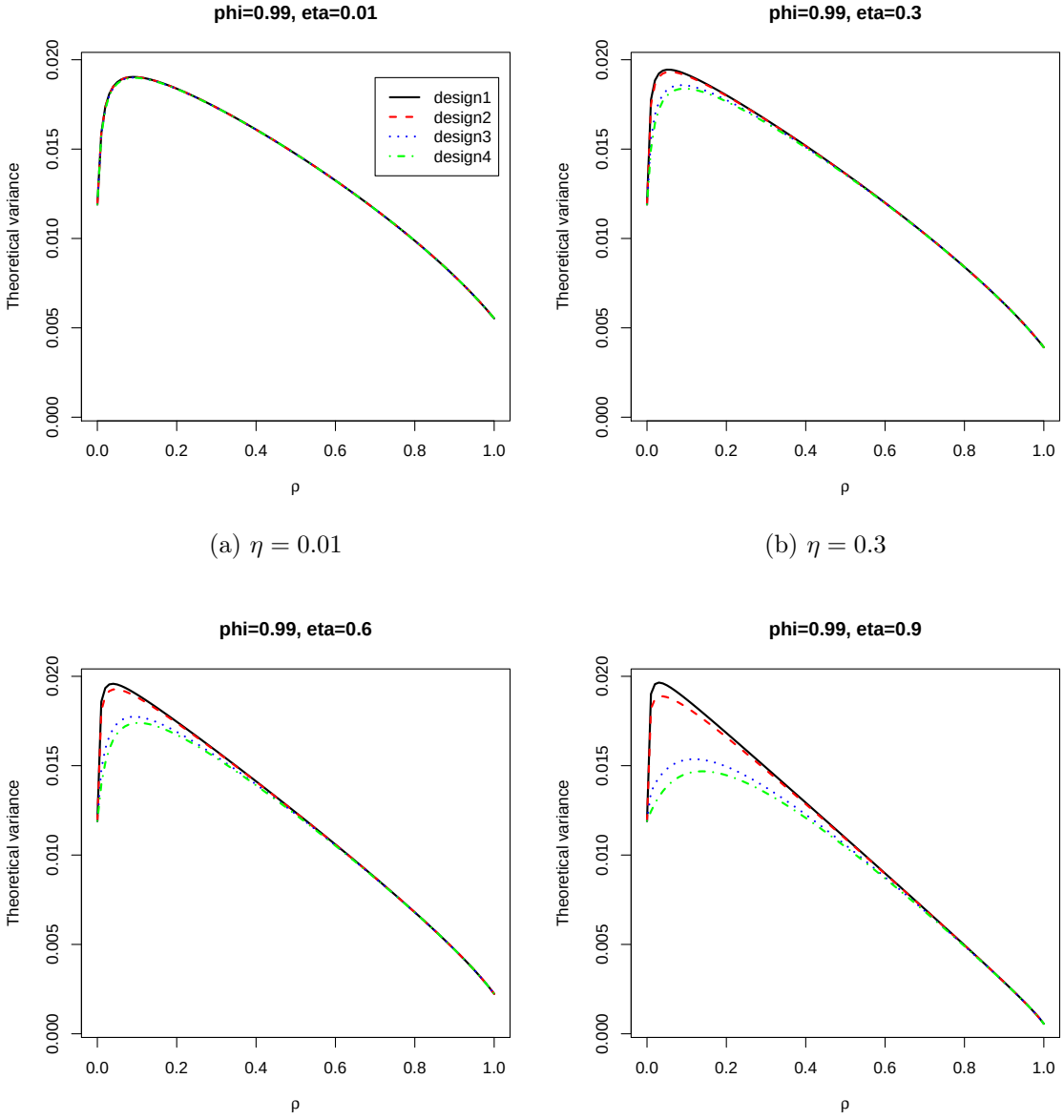


(a) $\eta = 0.01$ (b) $\eta = 0.3$



(c) $\eta = 0.6$ (d) $\eta = 0.9$

Figure 4.2: Var($\hat{\theta}$) vs. ρ ($\phi_b = 0.01$) when model (4.1) is correctly specified



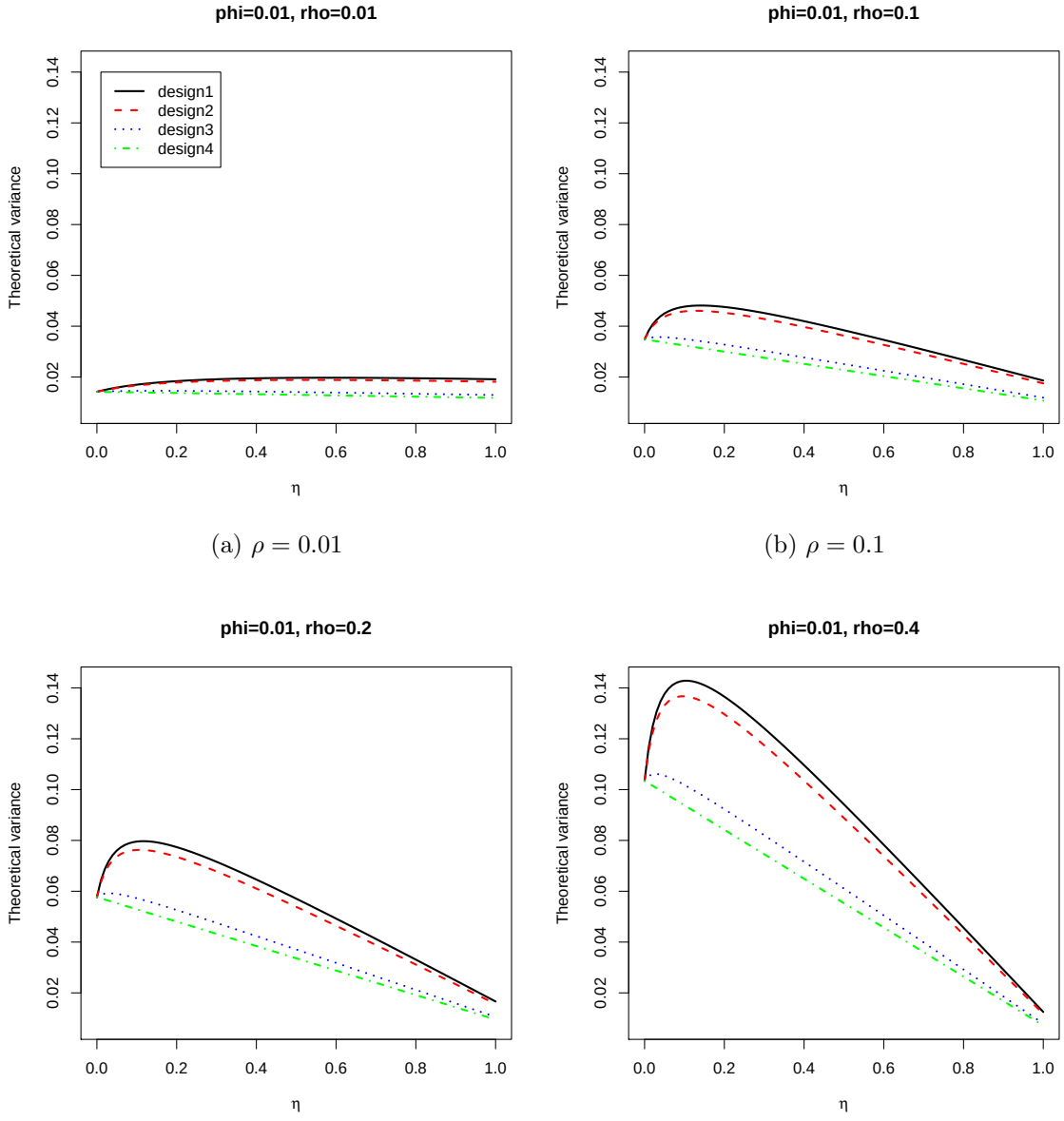
(a) $\eta = 0.01$

(b) $\eta = 0.3$

(c) $\eta = 0.6$

(d) $\eta = 0.9$

Figure 4.3: $\text{Var}(\hat{\theta})$ vs. ρ ($\phi_b = 0.99$) when model (4.1) is correctly specified



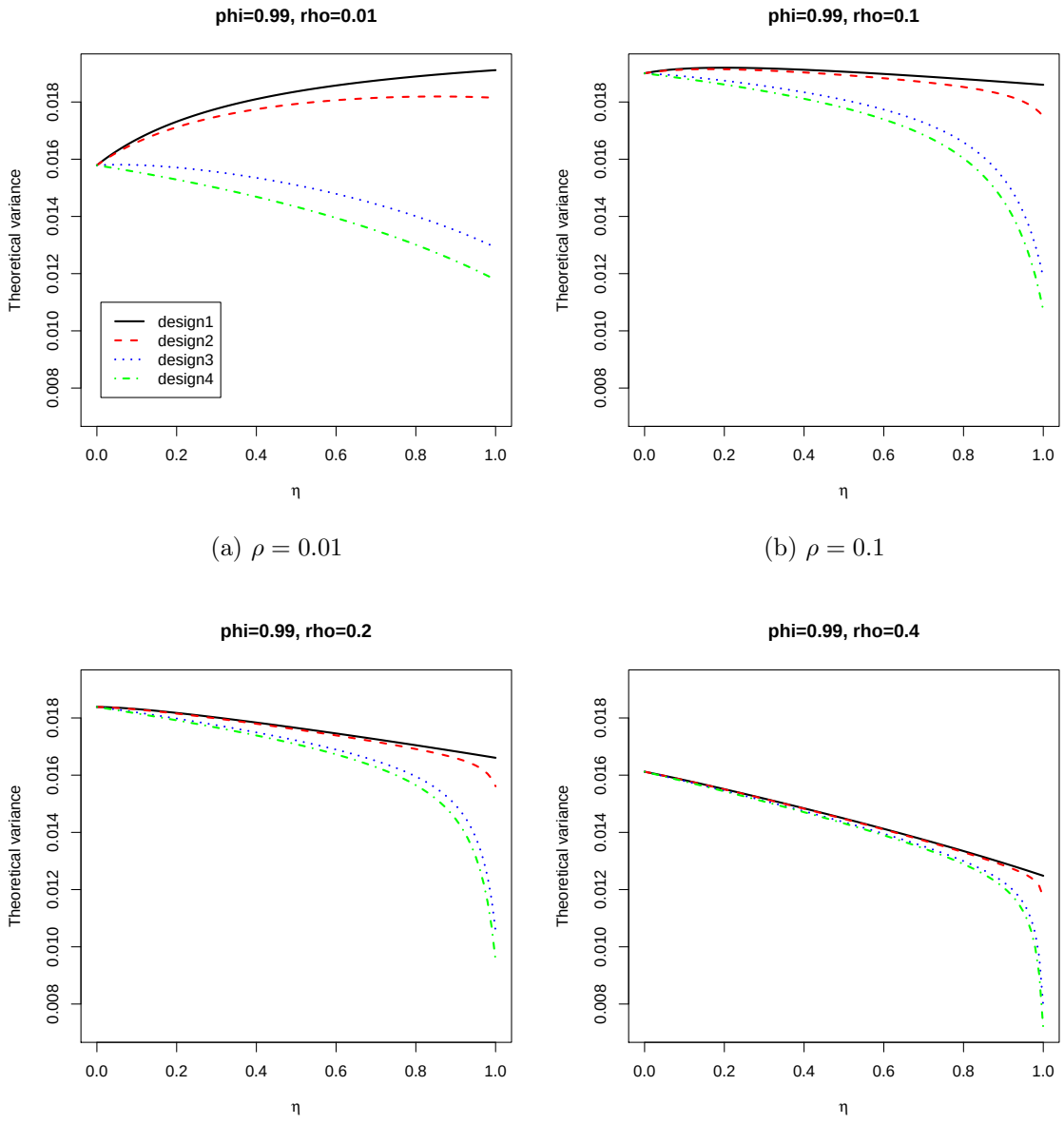
(a) $\rho = 0.01$

(b) $\rho = 0.1$

(c) $\rho = 0.2$

(d) $\rho = 0.4$

Figure 4.4: $\text{Var}(\hat{\theta})$ vs. η ($\phi_b = 0.01$) when model (4.1) is correctly specified



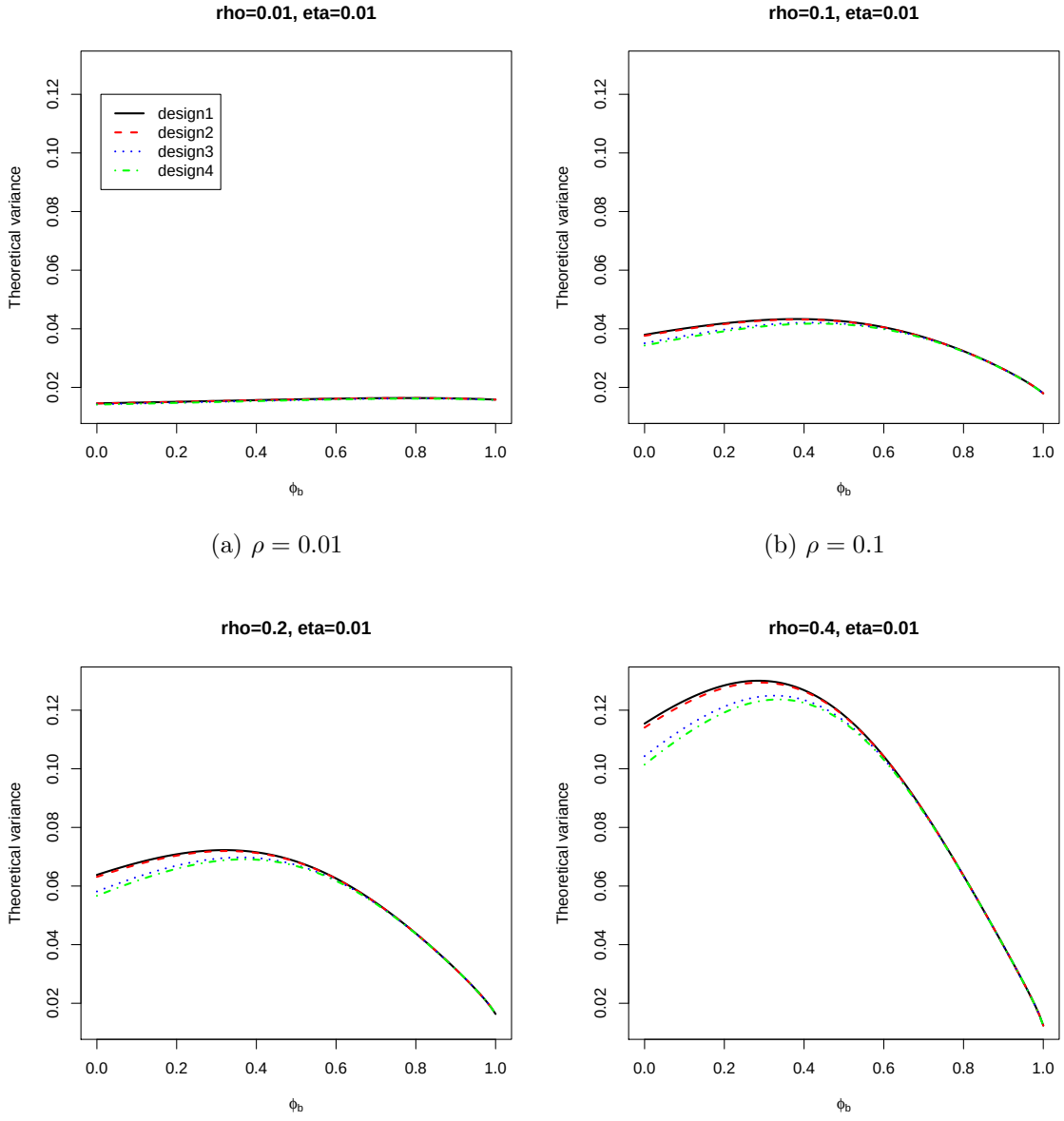
(a) $\rho = 0.01$

(b) $\rho = 0.1$

(c) $\rho = 0.2$

(d) $\rho = 0.4$

Figure 4.5: $\mathrm{Var}(\hat{\theta})$ vs. η ($\phi_b = 0.99$) when model (4.1) is correctly specified



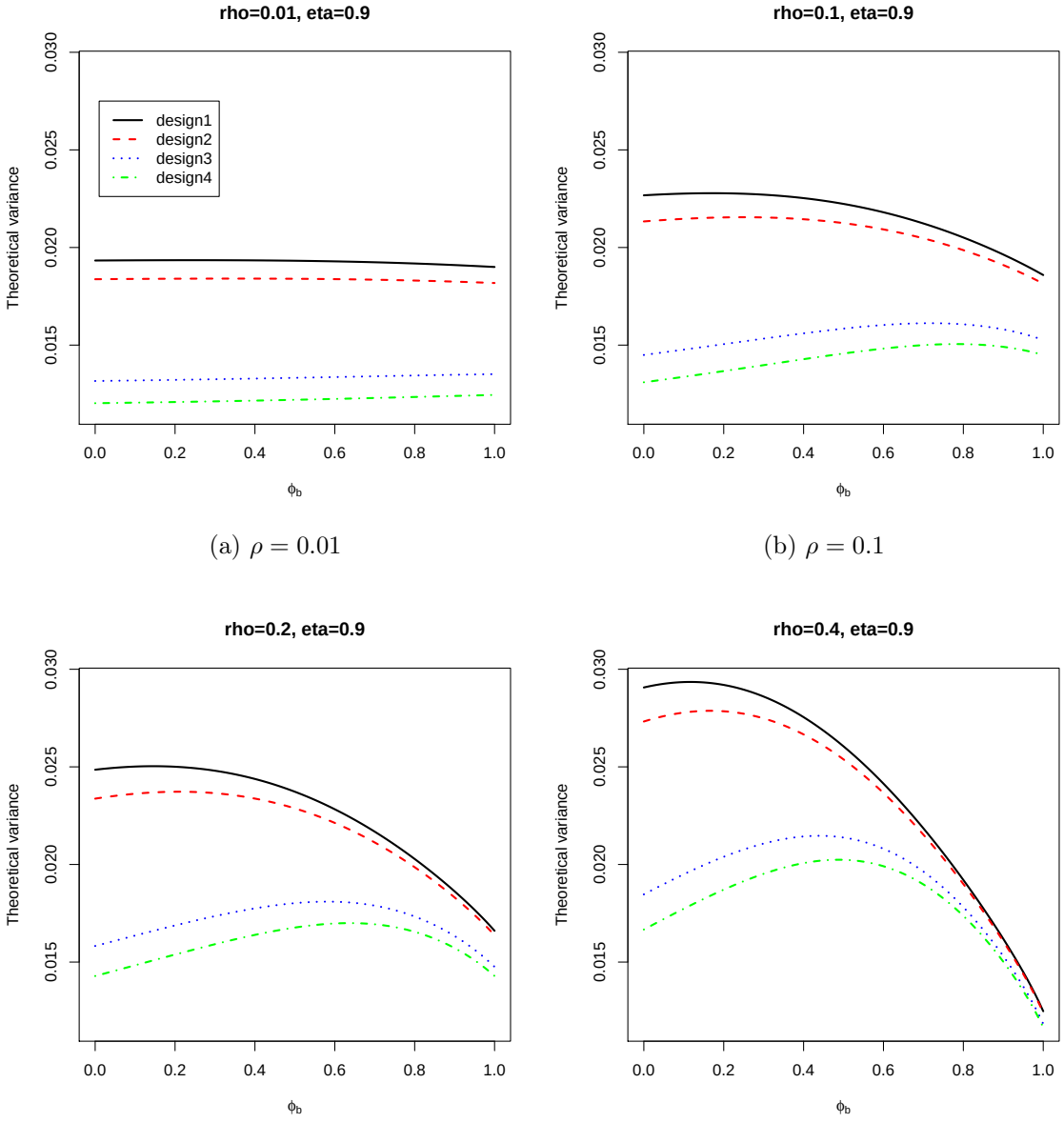
(a) $\rho = 0.01$

(b) $\rho = 0.1$

(c) $\rho = 0.2$

(d) $\rho = 0.4$

Figure 4.6: $\mathrm{Var}(\hat{\theta})$ vs. ϕ_b ($\eta = 0.01$) when model (4.1) is correctly specified



(a) $\rho = 0.01$

(b) $\rho = 0.1$

(c) $\rho = 0.2$

(d) $\rho = 0.4$

Figure 4.7: $\text{Var}(\hat{\theta})$ vs. ϕ_b ($\eta = 0.9$) when model (4.1) is correctly specified

4.2.2 Model Misspecification

In practice, the standard model (2.1) is frequently used to fit three-level data due to its simplicity. This use naturally raises the question about its performance in estimating the intervention effect θ when assumptions of this model are not met. We answer this question by examining the bias and coverage probability of the ML estimator of θ when fitting the standard three-level model (2.1) to data generated from model (4.1) that incorporates correlated random unit effects $b_{ijt}, t = 1, \cdots, T$ with autoregressive covariance pattern. Various scenarios are considered under each design.

Study Setup

We kept the same setup for $I, J, T, K, \delta_t (t = 1, \cdots, T), \theta$ and σ_y^2 as in Section 4.2.1 for all of the four designs. We set $\rho = 0.05$ and $\eta = 0.4$ to describe data collected from SW-CRTs. We varied the value of ϕ_b among $0.4, 0.6, 0.85, 0.95, 0.99$.

For each of the scenarios considered, we generated $10,000$ datasets from the underlying model (4.1) so that the Monte Carlo error allowed in estimating 95% confidence interval coverage probabilities is $\sqrt{0.95(1-0.95)/10000}$, approximately 0.0022. We then fit the data using the standard three-level model (2.1). For each dataset, we obtained the ML estimate as well as the 95% confidence interval for the treatment effect θ, and the ML estimates of the variance components $\sigma_a^2, \sigma_b^2, \sigma_e^2$. The coverage probability was computed by the proportion of the $10,000$ confidence intervals that include the true value of θ, 0.3. The model fitting of the three-level standard model was performed using R software package (R Core Team, 2016) with the `lme4` library (Bates et al., 2015).

Results

We show coverage probabilities for θ under all designs in Table 4.2. Empirical biases in $\hat{\theta}$ and estimators of the variance components $\sigma_a^2, \sigma_b^2, \sigma_e^2$ are given in Figure 4.8. Since the performance of the estimates of θ and the variance components are quite similar across the four designs, we only display them for design 1.

For a fixed design, the coverage probability for θ cannot even reach 0.9 when ϕ_b is as small as 0.4 and is close to the nominal coverage probability 0.95 as ϕ_b approaches 1 (Table 4.2). Given a small ϕ_b, the correlation between observations from the same unit at different times $\frac{\sigma_a^2+\sigma_b^2\phi_b^{|t-t'|}}{\sigma_a^2+\sigma_b^2+\sigma_e^2}, t \neq t'$ is not much stronger than the correlation between observations from different units within the same cluster $\frac{\sigma_a^2}{\sigma_a^2+\sigma_b^2+\sigma_e^2}$. However, under the standard model (2.1), the former correlation becomes $\frac{\hat{\sigma}_a^2+\hat{\sigma}_b^2}{\hat{\sigma}_a^2+\hat{\sigma}_b^2+\hat{\sigma}_e^2}$ and the latter correlation becomes $\frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2+\hat{\sigma}_b^2+\hat{\sigma}_e^2}$. Wrongly fitting the standard model would yield anticonservative results about θ. On the other hand, when ϕ_b is close to 1, the consequence of model misspecification is less serious due to the two models being almost equivalent. In addition, despite model misspecification, we note that the ML estimator $\hat{\theta}$ of the treatment effect is unbiased at all values of ϕ_b under all designs (Figure 4.8).

As for the estimates of the variance components σ_a^2, σ_b^2 and σ_e^2 under the incorrect model (2.1), they cannot provide unbiased estimates for the variance components under the correct model (4.1) (Figure 4.8). In fact, the interpretations of the variance components under these two models are different. Thus, it makes more sense to compare the estimated within-unit correlation under the incorrect model (2.1) to the underlying average within-unit correlation across time under the true model (4.1) for

a given ϕ_b. To compute the true average within-unit correlation under model (4.1), we followed the three steps as below (Corey et al., 1998).

1. For each within-unit correlation $\text{Corr}(Y_{ijtk}, Y_{ij(t+l)k'}) = \frac{\sigma_a^2+\sigma_b^2\phi_b^{|t-t'|}}{\sigma_a^2+\sigma_b^2+\sigma_e^2}, t = 1, \cdots, T; l = 0, \cdots, T-1$, we computed its Fisher's z-transformation (Fisher, 1958)

$$z_l = \frac{1}{2}\ln(\frac{1+\text{Corr}(Y_{ijtk}, Y_{ij(t+l)k'})}{1-\text{Corr}(Y_{ijtk}, Y_{ij(t+l)k'})}). \tag{4.4}$$

2. Compute the weighted average of these Fisher's z's transformed from the within-unit correlations under model (4.1).

$$\bar{z} = \frac{\sum_{l=0}^{T-1} w_l z_l}{\sum_{l=0}^{T-1} w_l},$$

where w_l takes the value of $(T-l)K^2$ when $l = 1, \cdots, T-1$ and $(TK^2-TK)/2$ when $l = 0$. The weight was chosen based on the number of repeated within-unit correlations for a given time difference l in the within-unit correlation matrix.

3. Using (4.4), transform the average Fisher's z back on the correlation scale to obtain the average within-unit correlation

$$\overline{\text{Corr}(Y_{ijtk}, Y_{ij(t+l)k'})} = \frac{e^{2\bar{z}}-1}{e^{2\bar{z}}+1}.$$

To obtain the estimated within-unit correlation under the incorrect model (2.1), we computed $\frac{\hat{\sigma}_a^2+\hat{\sigma}_b^2}{\hat{\sigma}_a^2+\hat{\sigma}_b^2+\hat{\sigma}_e^2}$ for each of the $10,000$ dataset, then computed their average. For a fixed ϕ_b, the true and the estimated average within-unit correlations are close (Table 4.3). Hence, if the standard three-level model (2.1) is used to fit data generated from model (4.1) with varying random unit effects across time, we tend to obtain an estimate of overall within-unit correlation across time.

Table 4.2: Coverage probabilities for θ given different values of ϕ_b under designs 1–4 when data from model (4.1) is misspecified to be from the standard model (2.1)

	ϕ_b				
	0.4	0.6	0.85	0.95	0.99
Design 1	0.8748	0.8898	0.9239	0.9406	0.9422
Design 2	0.8798	0.8938	0.9257	0.9371	0.9456
Design 3	0.882	0.9017	0.9209	0.9341	0.9435
Design 4	0.8868	0.8965	0.921	0.9418	0.9459

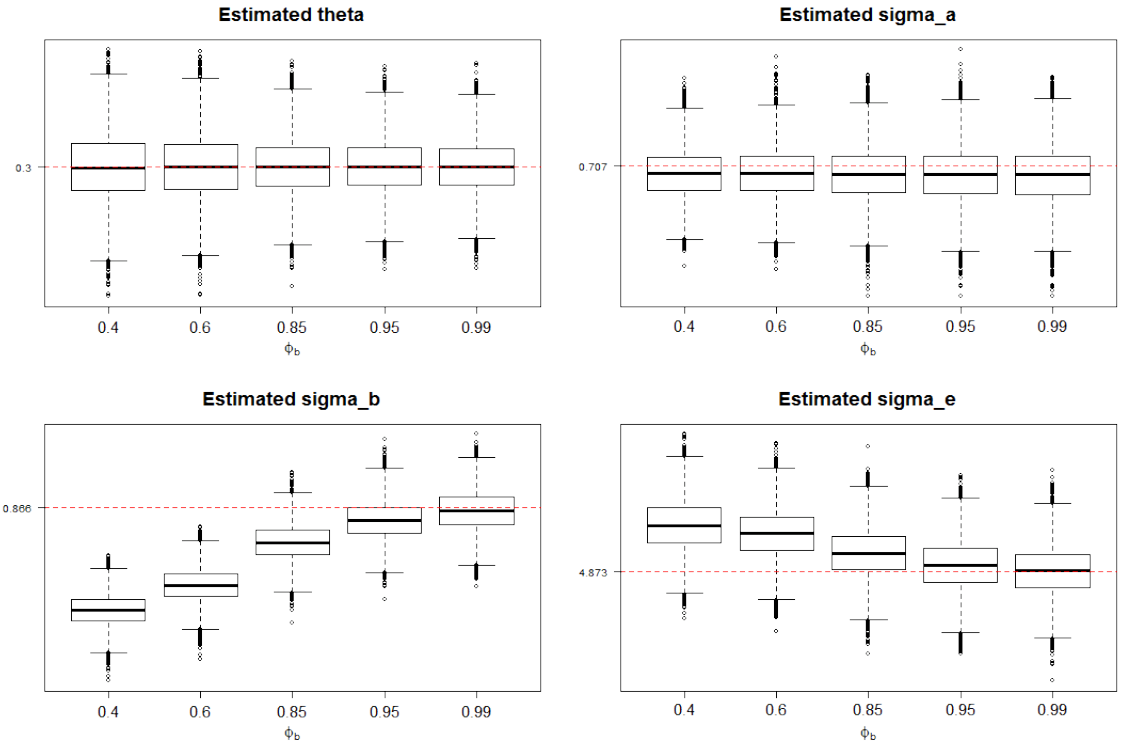


Figure 4.8: Boxplots of $\hat{\theta}$, $\hat{\sigma}_a$, $\hat{\sigma}_b$, and $\hat{\sigma}_e$ from 10000 repetitions under design 1 when data from model (4.1) is misspecified to be from the standard model (2.1). Red dashed lines denote true values of $\theta, \sigma_a, \sigma_b$ and σ_e, respectively.

Table 4.3: Average within-unit correlations under the true model (4.1) and the incorrectly fitted model (2.1) with the Autoregressive covariance pattern for random unit effects across time

ϕ_b	True average within-unit correlation	Estimated within-unit correlation
0.4	0.0285	0.0271
0.6	0.0326	0.0313
0.85	0.0414	0.0399
0.95	0.0468	0.0453
0.99	0.0493	0.0477

4.3 Correlated random unit effects with the Toeplitz Covariance Pattern

In this section, we generalize to the Toeplitz covariance pattern for pairs of random unit effects $(b_{ijt}, b_{ijt'}), t \neq t'$ under model (4.1). The relations between the theoretical variance of the intervention effect estimator and model parameters are examined in Section 4.3.1 when model (4.1) is correctly specified. In Section 4.3.2, we discuss consequences of fitting the simplified standard three-level model (2.1) to data generated from the underlying model (4.1).

4.3.1 Variance of the Intervention Effect Estimator under the Correct Model

Under the autoregressive covariance pattern (4.3) of $(b_{ijt}, b_{ijt'}), t \neq t'$, we showed that the theoretical variance of the GLS estimator of θ is affected by different values of ρ, η and ϕ_b when model (4.1) is correctly specified. Similarly, for model (4.1) with the Toeplitz covariance pattern (4.2) of $(b_{ijt}, b_{ijt'}), t \neq t'$, we are interested in how ρ, η and different speeds and ranges of decay in $\text{Corr}(b_{ijt}, b_{ijt'})$ affect the theoretical

variance of $\hat{\theta}$ when the model is correctly specified. The above question of interest will be discussed under designs 1–4.

Study Setup

In order to generate Toeplitz correlation matrices, we considered a typical type of hub correlations where the correlations of the 1st observation and all the other observations are known, and the correlation between the 1st observation and the jth observation decays as j increases (Hardin et al., 2013). Suppose that the first row of a $T \times T$ correlation matrix P consists of the following values

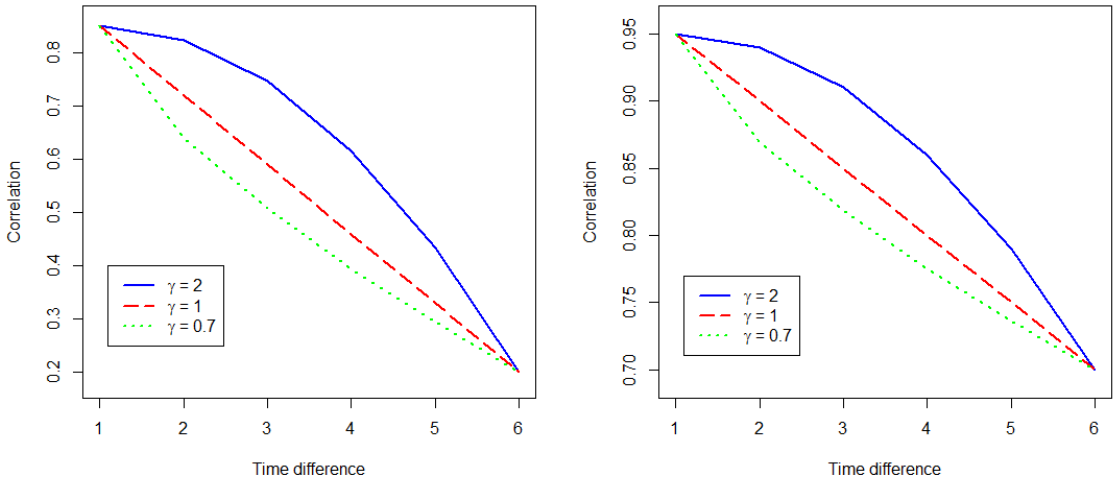
$$P_{11} = 1, P_{1j} = \rho_{max} - (\rho_{max} - \rho_{min})(\frac{j-2}{T-2})^{\gamma}, \tag{4.5}$$

where $j = 2, \cdots, T$ and P_{1j} is the $(1, j)$th entry of matrix P. Note that with this specification, P_{1j} decreases from ρ_{max} to ρ_{min} as j increases from 2 to T. Furthermore, the parameter γ is a tuning parameter for the speed of decay in the correlations. In the special case of $\gamma = 1$, the correlations decay in a linear pattern. Based on the hub correlations, Toeplitz correlation matrices can be formed.

We considered two possible ranges of decay in $\text{Corr}(b_{ijt}, b_{ijt'})$ as $|t - t'|$ increases. First, we set ρ_{max} to be 0.85 and ρ_{min} to be 0.2. Although unlikely in reality, this large range can provide insights about the performance of $\text{Var}(\hat{\theta})$ when model (4.1) is correctly specified. Second, we considered a more realistic scenario and set ρ_{max} to be 0.95 and ρ_{min} to be 0.7. Given a fixed range of decay, the speed of decay can be different. In light of this, we chose γ to be $2, 1, 0.7$, which yields quadratic, linear and 0.7-degree decay, respectively. Figure 4.9 displays different decay speeds when decay ranges are large and small, respectively. Note that the 0.7-degree decay speed

is similar to the decay speed in the autoregressive covariance pattern where the speed is fast initially and slows down gradually.

We kept the same setup for $I, J, T, K, \delta_t (t = 1, \cdots, T), \theta$ and σ_y^2 as in Section 4.2.1 for all of the four designs. For fixed range and speed of decay in $\text{Corr}(b_{ijt}, b_{ijt'})$ as $|t - t'|$ increases, the ICC ρ varies among $0.01, 0.1, 0.2$ when we studied the relations between $\text{Var}(\hat{\theta})$ and η, and η varies among $0.1, 0.5, 0.9$ when we studied the relations between $\text{Var}(\hat{\theta})$ and ρ. Lastly, the GLS estimator of θ was obtained by computing $(X^T V^{-1} X)^{-1}$, where X and V are the design matrix and the covariance matrix of the outcome under model (4.1), respectively.



(a) Large decay range

(b) Small decay range

Figure 4.9: Different speeds of decay in $\text{Corr}(b_{ijt}, b_{ijt'})$ as time difference $|t - t'|$ increases.

Results

We consider the case of a large decay range, increasing ρ and fixed η (Figure 4.10) to explain how different speeds of decay in $\text{Corr}(b_{ijt}, b_{ijt'}), t \neq t'$, affect the performance of $\text{Var}(\hat{\theta})$. When η is fixed at 0.1 (left column in Figure 4.10), the patterns in the theoretical variance are substantially different under designs 1–4 depending on the speed of decay in $\text{Corr}(b_{ijt}, b_{ijt'}), t \neq t'$. For the quadratic decay ($\gamma = 2$), the theoretical variance first increases and then decreases under all designs with the gap between designs 1, 2 and designs 3, 4 widening as ρ increases. For the linear decay ($\gamma = 1$), the theoretical variance monotonically increases under all designs and there are minimal differences among the designs. Lastly, for the 0.7-degree decay ($\gamma = 0.7$), the theoretical variance first increases when ρ increases from 0 to 0.8 and then has a slight decrease as ρ approaches 1. Same with the case of linear decay, the differences among the four designs are minimal.

Revisiting the three types of comparisons discussed in Section 4.2, we have two competing sets of comparisons – within-cluster comparisons (types 1 and 2) and between-cluster comparisons (type 3) under model (4.1). When ρ increases and η is fixed, the effects of within-cluster comparisons are improving and the effects of between-cluster comparisons are worsening. For the quadratic decay ($\gamma = 2$), the worsening effects of between-cluster comparisons are dominating at small values of ρ. After ρ passes through some threshold (around 0.2), the improving effects of within-cluster comparisons are dominating and thus increase the precision. Meanwhile, the effect of between-unit within-time comparisons (included in type 2 comparisons) is most beneficial to designs 3 and 4, which leads to different theoretical variances among the four designs. Following a similar argument, we can explain the patterns

of the theoretical variance for the case of 0.7-degree decay. However, in this case the threshold of ρ is much larger and close to 0.8. Also, the effect of between-unit within-time comparisons does not seem to play an important role in improving precision, since all designs yield minimal difference in theoretical variance of $\hat{\theta}$. Lastly, under the linear decay ($\gamma = 1$) the worsening effects of between-cluster comparisons are dominating at all values of ρ, which leads to monotonically increasing variance under all designs. Although less sharp, the contrast in theoretical variance between different decay speeds when η is fixed at a larger value can still be easily detected (right two column in Figure 4.10).

When the decay range is small and η is fixed at 0.1, different patterns in the theoretical variance are again observed for the three decay speeds (left column in Figure 4.11). In this scenario, the maximum correlation between $(b_{ijt}, b_{ijt'}), t \neq t'$ is 0.95 and the minimum correlation is 0.7. These high correlations contribute to both improving effects of within-cluster comparisons (types 1 and 2) and worsening effects of between-cluster comparisons (type 3). Depending on the magnitude of ρ, either one of the two competing sets of comparisons would dominate the other. Under each decay speed and each design, the theoretical variance of $\hat{\theta}$ increases at small values of ρ and decreases when ρ exceeds some threshold. When γ is fixed at $2, 1$ and 0.7, the threshold of ρ is around $0.2, 0.6$ and 0.5, respectively, under all designs. As observed, when the range of decay in the correlation between $\text{Corr}(b_{ijt}, b_{ijt'}), t \neq t'$ is small, the patterns in $\text{Var}(\hat{\theta})$ are not as dramatically different as when the range is large across the three decay speeds.

When ρ is fixed and η is increasing, fanning patterns among the four designs are observed regardless of the range or speed of decay in $\text{Corr}(b_{ijt}, b_{ijt'}), t \neq t'$ (Figures

4.12 and 4.13). The gap between design 1, 2 and designs 3, 4 increases with η. Given constant ρ, between-unit within-cluster comparisons (type 2) have improving effects and between-cluster comparisons (type 3) have worsening effects on the precision of $\hat{\theta}$ when η increases. For example, given the small decay range and small ρ (0.01), between-cluster comparisons and within-cluster comparisons dominate at all values of η under designs 1, 2 and 3, 4, respectively. This dominance holds for all considered decay speeds. The decay speeds, on the other hand, mainly affect the overall magnitude of the variance of $\hat{\theta}$ under each design. Compared to the case of increasing ρ for fixed η (Figures 4.10 and 4.11), different ranges and speeds of decay in the correlation of $\text{Corr}(b_{ijt}, b_{ijt'}), t \neq t'$ do not yield sharp contrast when increasing η for fixed ρ.

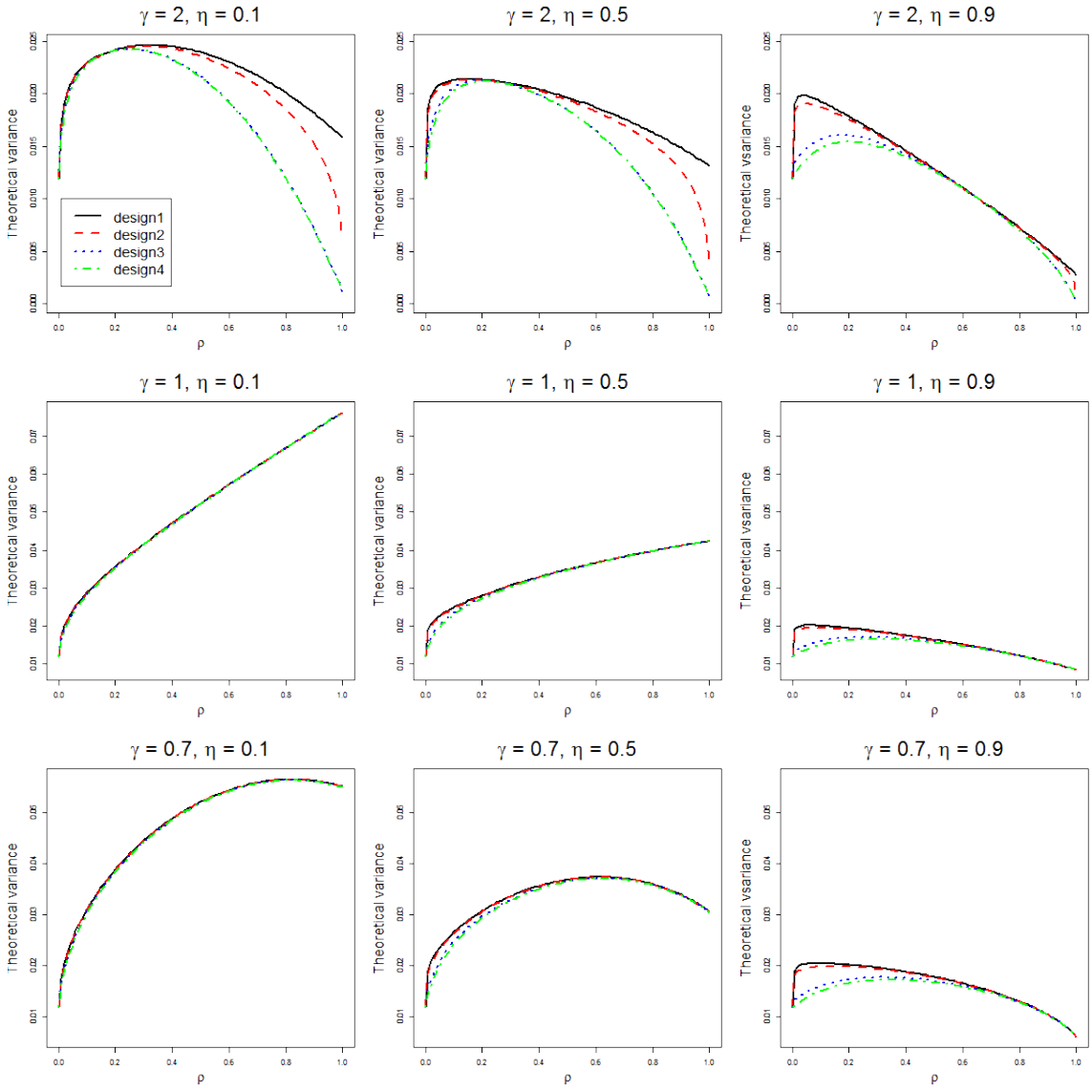


Figure 4.10: $\mathrm{Var}(\hat{\theta})$ vs. ρ given the large decay range when model (4.1) is correctly specified.

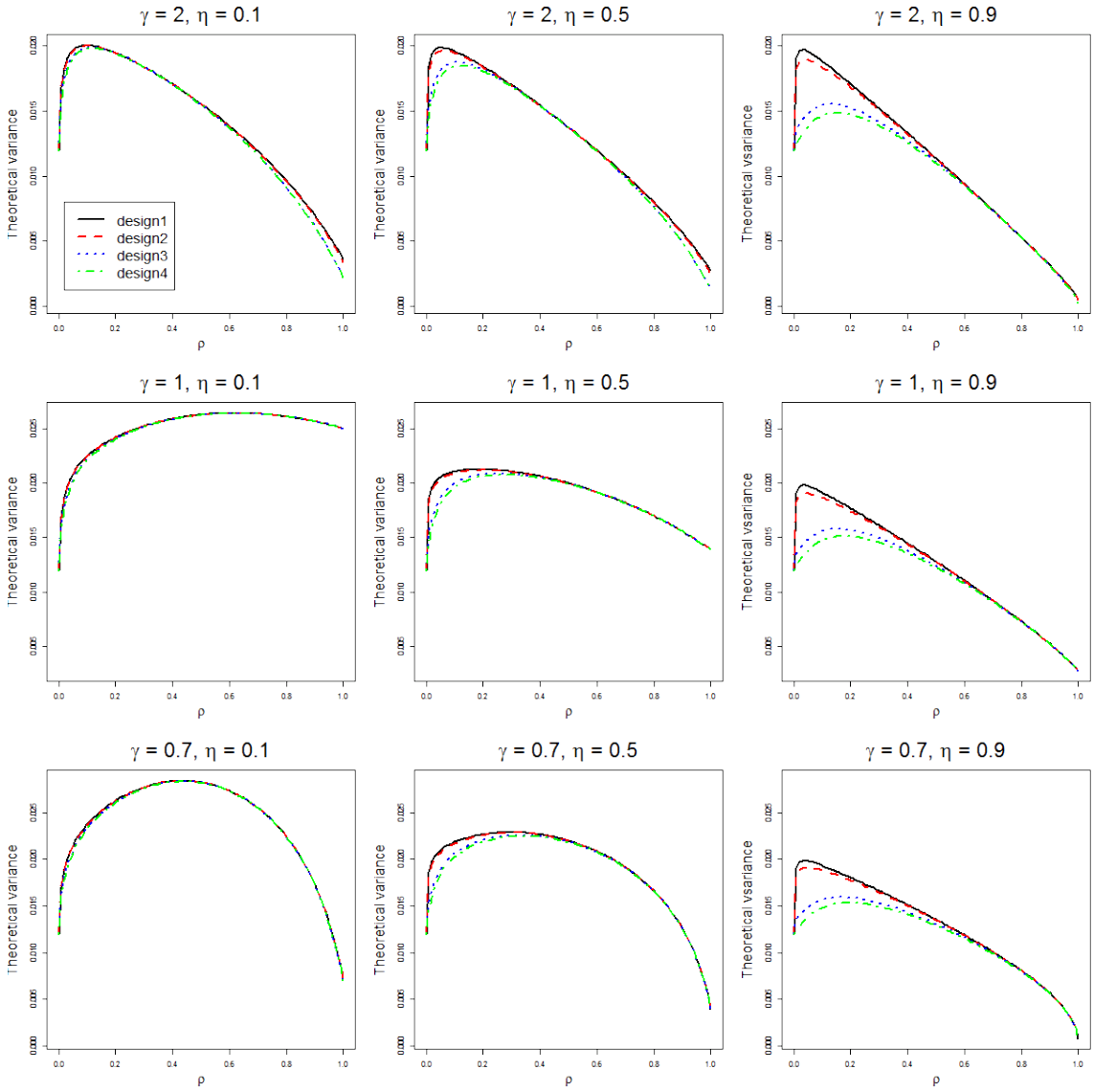


Figure 4.11: Var($\hat{\theta}$) vs. ρ given the small decay range when model (4.1) is correctly specified.

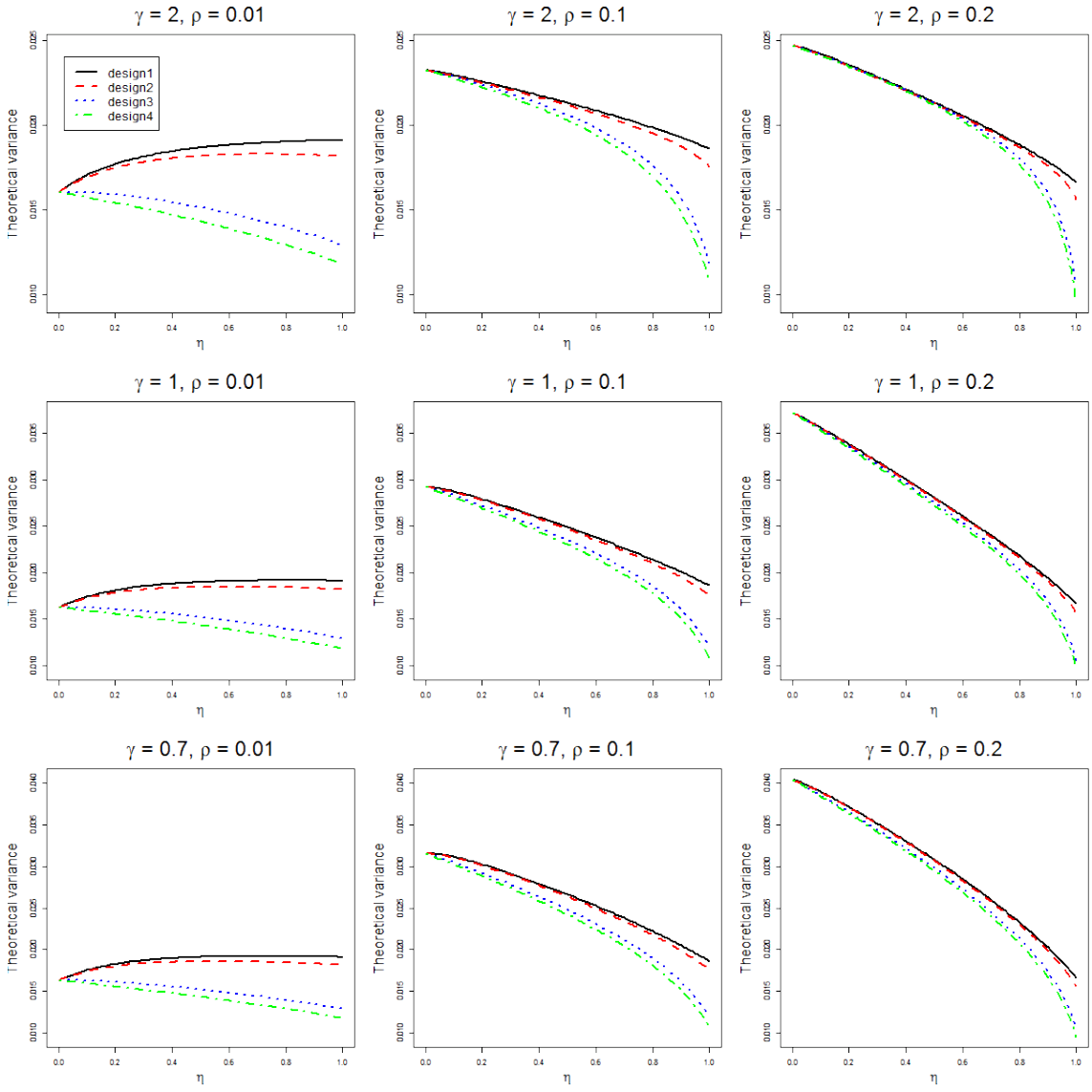


Figure 4.12: Var($\hat{\theta}$) vs. η given the large decay range when model (4.1) is correctly specified.

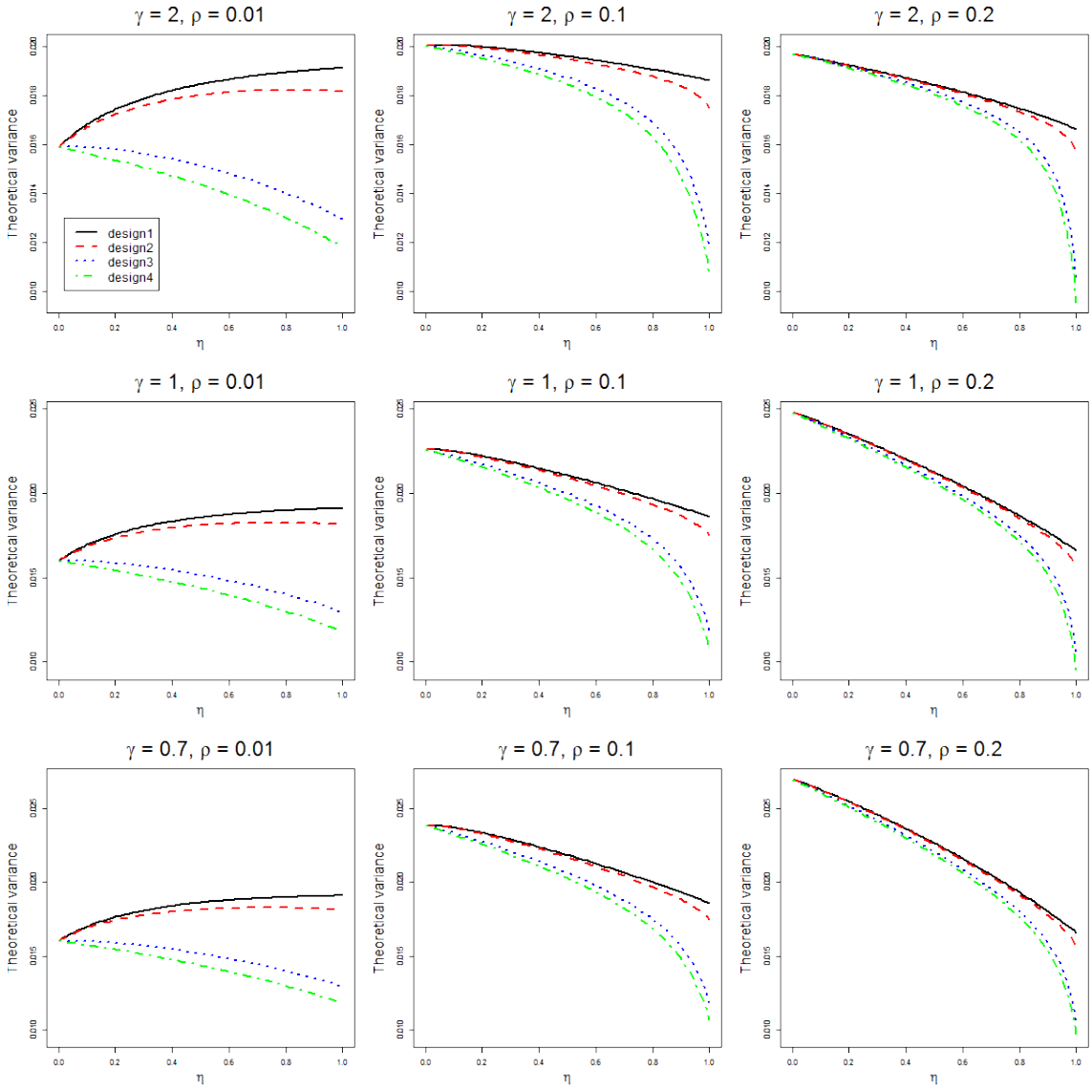


Figure 4.13: $\mathrm{Var}(\hat{\theta})$ vs. η given the small decay range when model (4.1) is correctly specified.

4.3.2 Model Misspecification

As in Section 4.2.2, we study the bias and coverage probability of the ML estimator of θ when the standard three-level model (2.1) is fit to data generated from model (4.1) incorporating correlated random unit effects $b_{ijt}, t = 1, \cdots, T$. The effect of range and speed of decay in $\text{Corr}(b_{ijt}, b_{ijt'}), t \neq t'$ on model misspecification is considered. We also discuss the performance of the estimators of the variance components σ_a^2, σ_b^2 and σ_e^2.

Study Setup

We followed from the setup for the ranges and speeds of decay in $\text{Corr}(b_{ijt}, b_{ijt'}), t \neq t'$ in our previous subsection, and kept the same setup for $I, J, T, K, \delta_t (t = 1, \cdots, T), \theta$ and σ_y^2 as in Section 4.2.1 for all four designs. In addition, the ICC ρ was fixed at 0.05, and the ratio of between-unit to within-unit correlations η at 0.4.

For each of the scenarios considered, we generated $10,000$ datasets from the underlying model (4.1) so that the Monte Carlo error allowed in estimating 95% confidence interval coverage probabilities is $\sqrt{0.95(1-0.95)/10000}$, approximately 0.0022. We then fit each dataset using the standard three-level model (2.1). For each dataset, we obtained the ML estimate of the treatment effect θ, the ML estimate of the variance components $\sigma_a^2, \sigma_b^2, \sigma_e^2$, and a 95% confidence interval for θ. The coverage probability was computed by the proportion of the confidence intervals that include the true value of θ, 0.3. The model fitting of the three-level standard model was performed using R software package (R Core Team, 2016) with the `lme4` library (Bates et al., 2015).

Results

We show coverage probabilities for θ under all designs in Table 4.4. Empirical biases in $\hat{\theta}$ and estimators of the variance components $\sigma_a^2, \sigma_b^2, \sigma_e^2$ are given in Figure 4.14. Since the performance of the estimates of θ and the variance components are quite similar across the four designs, we only display them for design 1.

Under all designs, empirical coverage probabilities of the 95% confidence interval for θ are lower than the nominal coverage probability (Table 4.4). The highest coverage probabilities occur in the case of small decay range and quadratic decay speed ($\gamma = 2$). In addition, for the same design and same decay range, the quadratic decay speed leads to the highest coverage probability among the three decay speeds. The table also shows that for the same design and same decay speed, the small decay range yields higher coverage probability than the large decay range. Hence, the coverage probability is close to the pursued level 0.95 if the correlations between $(b_{ijt}, b_{ijt'}), t \neq t'$ are on average close to the assumed correlation 1 under the standard three-level model.

When data are generated from model (4.1) with the Toeplitz covariance pattern in $\text{Corr}(b_{ijt}, b_{ijt'}), t \neq t'$, fitting the standard model (2.1) does not bias the ML estimator of the intervention effect θ regardless of the speed or range of decay in the Toeplitz covariance matrix (Figure 4.14).

Figure 4.14 also shows that fitting the incorrect model (2.1) cannot provide unbiased estimates of the variance components σ_a^2, σ_b^2 and σ_e^2 under the correct model (4.1), as variance components in these two models have different interpretations. Hence, we compared the estimated within-unit correlation under the incorrect model and the underlying average within-unit correlation across time for a given speed and decay

of $\text{Corr}(b_{ijt}, b_{ijt'}), t \neq t'$. We followed the same methods as in Section 4.2.2 to compute the true and the estimated average within-unit correlations. For a fixed decay range and speed, the true and the estimated average within-unit correlations are close (Table 4.5). Hence, when we fit the incorrect model (2.1) assuming constant within-unit correlation to data generated from model (4.1) with varying random unit effects across time, we may actually obtain an estimate of the overall within-unit correlation across time.

Table 4.4: Coverage probabilities under designs 1–4 and two possible types of decay of correlations

	Large decay range			Small decay range		
	$\gamma = 2$	$\gamma = 1$	$\gamma = 0.7$	$\gamma = 2$	$\gamma = 1$	$\gamma = 0.7$
Design 1	0.9308	0.9195	0.9156	0.9383	0.9402	0.9386
Design 2	0.9365	0.92	0.9174	0.9457	0.9403	0.9392
Design 3	0.9317	0.9218	0.9111	0.9418	0.9391	0.9372
Design 4	0.9318	0.9184	0.9098	0.9438	0.9383	0.9377

Table 4.5: Average within-unit correlations under the true model (4.1) and the incorrectly fitted model (2.1) with the Toeplitz covariance pattern for random unit effects across time

Decay range	Decay speed	True average within-unit correlation	Estimated within-unit correlation
Large	Quadratic	0.0427	0.0413
	Linear	0.0405	0.0391
	0.7-degree	0.0393	0.0378
Small	Quadratic	0.0474	0.0459
	Linear	0.0465	0.0451
	0.7-degree	0.0461	0.0445

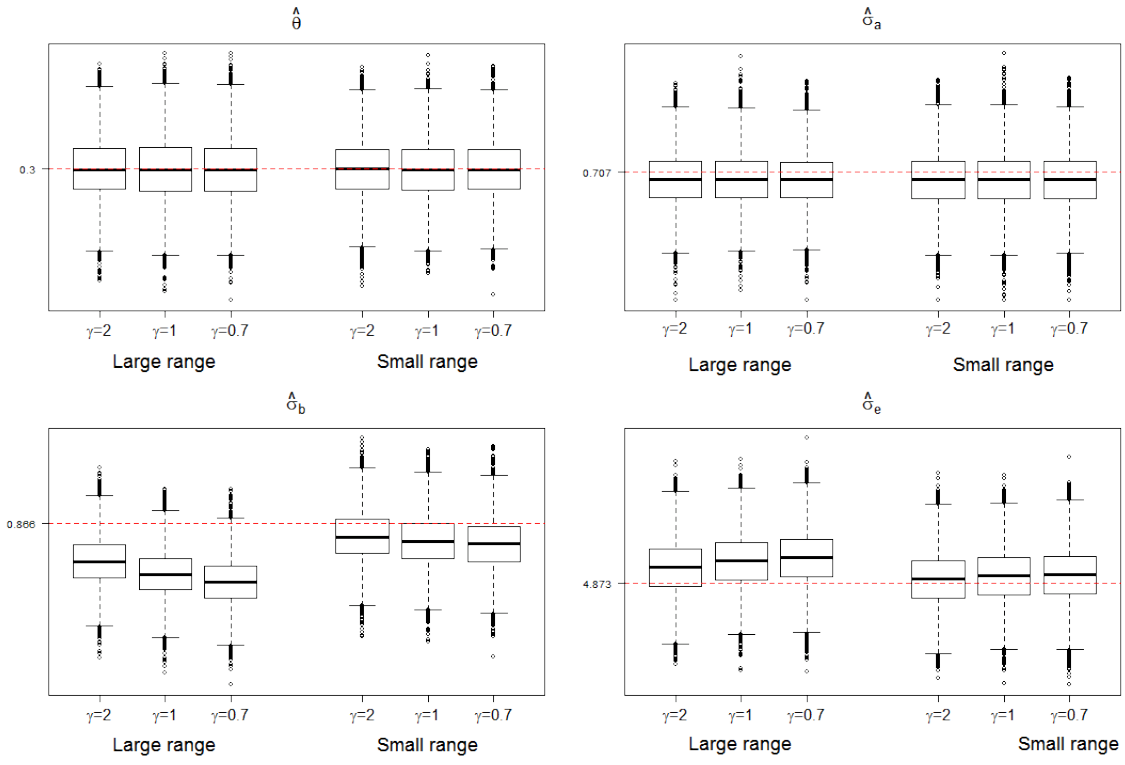


Figure 4.14: Boxplots of $\hat{\theta}$, $\hat{\sigma}_a$, $\hat{\sigma}_b$, and $\hat{\sigma}_e$ from 10000 repetitions under design 1 when data from model (4.1) is misspecified to be from the standard model (2.1). Red dashed lines denote true values of $\theta, \sigma_a, \sigma_b$ and σ_e, respectively.

4.4 Discussion

In this chapter, we extended the three-level standard model (2.1) to a model with changing unit effects across time points to generate data from three-level SW-CRTs . Unlike the standard model where the random unit effect b_{ij} is identical across all time points, the proposed model incorporates varying random unit effects $b_{ijt}, t = 1, \cdots, T$ for T time points. While it provides practical usefulness, it is unclear how the model parameters impact the precision of the ML estimator of the intervention effect θ. We considered the autoregressive and the general Toeplitz covariance patterns as typical

examples to model the correlation between the pairs $(b_{ijt}, b_{ijt'}), t \neq t', t = 1, \cdots T$. By referring to the three types of comparisons, we thoroughly studied the relations between the variance of $\hat{\theta}$ and the parameters ρ, η, ϕ_b (under the AR covariance pattern), and $\phi_{b,t}, t = 1, \cdots, T-1$, (under the Toeplitz covariance pattern).

Our results for model misspecification cast some doubt on the popular practice of using the standard three-level model for three-level SW-CRTs without considering possibly different random unit effects over time. Both Section 4.2 and Section 4.3 show that when the standard model (2.1) is fit to data generated from the underlying model (4.1), the ML estimator of the intervention effect θ is still unbiased. However, such model misspecification would lower the empirical coverage probabilities for the intervention effect. We also note that when incorrectly fitting the standard model assuming constant within-unit correlation across time to data generated from the underlying model (4.1) assuming varying within-unit correlation across time, we are actually estimating the average within-unit correlation across time. In practice, if investigators believe that unit effects are likely to change with time, then models like (4.1) including varying random unit effects may be preferred over the standard model (2.1) with constant random unit effect.

Chapter 5: Contamination under Three-level SW-CRTs

When randomization is conducted at the individual level for nested data, a key issue that may arise is control group contamination as information may leak from the intervention group to the control group (Donner and Klar, 1994). Analogously, our proposed three-level SW-CRTs randomizing units within clusters (designs 2–4) may yield the same contamination problem for participants assigned to the control units. Using the example of patients within wards (units) within hospitals (clusters), if patients from intervention wards share information about the intervention with patients from control wards, control wards may be under the risk of contamination.

Contamination is commonly described in terms of two aspects, rate and intensity (Keogh-Brown et al., 2007). Contamination rate is defined to be the proportion of subjects in the control arm being exposed to the intervention. Contamination intensity is the proportion of the intervention effect on participants who are contaminated. Typically, if the underlying effect of the intervention on subjects in the intervention arm is θ, then the effect of the intervention on subjects from the control arm is $c \cdot \theta$ where $c \in [0, 1]$ indicates the intensity of contamination.

Throughout this chapter, we assume that no contamination can happen under design 1. Following the scheme of design 1, the units of randomization are clusters. Within each cluster, all participating units follow the same assignment and receive

either the control or the intervention at the same time. Hence, units are assumed to have minimal risk of contamination, and clusters are assumed to share little information about the intervention with one another. For example, hospitals may not be able to share information due to long geographical distance.

In this chapter, we consider two scenarios where contamination could occur in SW-CRTs. In Section 5.1, we introduce how contamination can occur under each of designs 2–4. Consequences of contamination are then discussed for each design. In Section 5.2, we consider a crossed allocation of nurses to wards within a hospital. Under each design, we study the impact of contamination of control wards due to shared nurses with intervention wards, and the impact of ignoring the nurse effect in modeling data from trials with randomization being conducted at ward level. Section 5.3 includes a discussion of this chapter.

5.1 Contamination and Model Misspecification for Assuming Ward Effects When None Are Present

We consider a scenario where contamination happens between units assigned to two different treatments within the same cluster due to highly shared treatment resources. For example, patients from different wards may share many of the same doctors or nurses. This shared medical resource may lead to correlated patients from different wards and minimal difference among patients from different wards. In some cases, it may even lead to contamination when the same doctor or nurse simultaneously gives new care to intervention patients and standard care to control patients. In this case, contamination may happen to control patients if their doctor or nurse changes behavior based on the intervention. In addition, minimal variation among

different wards may lead to model misspecification by wrongly assuming a ward random effect.

In this section, we consider the above scenario and investigate the consequence of contamination induced by allocation of wards within the same hospital to different treatments. Severity of contamination of control wards is discussed under different stepped wedge designs. We use numerical examples to study the impact of ignoring contamination on inference about the intervention effect. We also show results of fitting models that incorrectly include a ward effect while the underlying model to generate the data from these trials does not contain the ward random effect.

5.1.1 The Model

When the unit effect does not exist in data collected from three-level CRTs, the standard three-level model reduces to a two level model where the random unit effect is eliminated. Consider the following two-level model

$$Y_{ijtk} = \delta_t + \theta Q_{ijtk} + a_i + \epsilon_{ijtk}, \tag{5.1}$$

where $i = 1, \cdots, I$ is the index for independent clusters, $j = 1, \cdots, J$ is the index for units nested within the same cluster, $t = 1, \cdots, T$ is the index for time points, and $k = 1, \cdots, K$ is the index for observations from the same unit within a cluster at the same time. Accordingly, Y_{ijtk} denotes the kth observation from the jth unit nested in the ith cluster at time t. The random cluster effects $a_i, i = 1, \cdots, I$ are independently drawn from $N(0, \sigma_a^2)$. The random errors ϵ_{ijtk} are independently drawn from $N(0, \sigma_e^2)$. Furthermore, the random effects a_i and the error ϵ_{ijtk} are mutually independent. The fixed effect of time t is denoted by δ_t. The intervention effect is denoted by θ.

The variable Q_{ijtk} is equal to 1 if patient k from ward j within hospital i is assigned to the intervention at time t, 0 if the patient is assigned to the control and not contaminated at time t, and a value between 0 and 1 if the patient is assigned to the control but subject to contamination at time t. A value of Q_{ijtk} between 0 and 1 represents the proportion of intervention received by the contaminated control participant. A large value of Q_{ijtk} indicates high contamination intensity.

5.1.2 Problem Setup and Contamination under Each Design

Under designs 2–4, contamination happens at different times and lasts for different lengths of time. Under design 2, one step after the first half of a cluster switches to the intervention, the second half switches to the intervention. In this case, contamination may be mild. Figure 5.1 provides visualization of contamination of control units in each cluster. In design 2, there are three different moments of possible contamination depending on when each cluster start and finish the crossover to the intervention. In clusters 1 and 2, units 4–6 may have contamination at time 2. In clusters 3 and 4, units 4–6 may have contamination at time 4. In clusters 5 and 6, units 4–6 may have contamination at time 6. Similarly, under design 3, units 4–6 in clusters 1 and 2 may be contaminated during time 2 and time 4, units 4–6 in clusters 3 and 4 may be contaminated during time 3 and 5, and units 4–6 in clusters 5 and 6 may be contaminated during time 4 and 6 (Figure 5.2). Lastly, under design 4 where a small-scale stepped wedge study is conducted within each cluster, units 2–6 in each cluster can potentially be exposed to the intervention during time 2, 2–3, 2–4, 2–5 and 2–6, respectively (Figure 5.3).

As before, we considered 18 independent clusters, 6 units within each cluster, and 20 subjects recruited in each unit at each time. The duration of each trial was 7 time periods, where subject data are collected at each time point. The total variation of the outcome was $\sigma_y = 5$. Under the two-level model (5.1), we set $\rho = 0.1$. The intervention effect θ was 0.4 so that the power of detecting $\theta = 0.4$ achieved 80% under model (5.1) with $c = 0$ for all control subjects. For ease of analysis, we assumed no time effects and thus $\delta_1 = \cdots = \delta_T = 50$.

We considered three possible contamination rates and five possible levels of contamination intensity. When a fixed proportion of control patients within a ward is contaminated, the intensity c varies among $0.2, 0.4, 0.6, 0.8$ and 1. When contamination intensity c is fixed, the rate varies among $0.3, 0.6$ and 1. A rate of 1 indicates all control patients from a ward being contaminated. In addition, we also considered the scenario of no contamination by setting either rate or intensity to 0.

For each of the scenarios considered, we generated $10,000$ datasets from the underlying model (5.1). In this case, the Monte Carlo error allowed in estimating 95% confidence interval coverage probabilities is $\sqrt{0.95(1-0.95)/10000}$, approximately 0.0022.

Our analysis includes four parts. First, we fit the true data-generating model (5.1) as a reference. Second, we fit a model that correctly incorporates contamination rate and intensity of control subjects, but wrongly assumes the existence of random unit effect. We used model

$$Y_{ijtk} = \delta_t + \theta Q_{ijtk} + a_i + b_{ij} + \epsilon_{ijtk},$$

where b_{ij} is random unit effect and other notations are the same as in the true model (5.1). Third, we fit a model that correctly specifies the random effects, but fails

to acknowledge contamination happened to control subjects. In other words, the standard two-level model (1.7) was used to fit the data. Lastly, we fit a model both misspecifying the random unit effect and disregarding contamination. In this case, the standard three-level model (2.1) was used to fit the data.

		Time						
Pattern	(Cluster, Units)	1	2	3	4	5	6	7
1	(1, 1-3); (2, 1-3)	0	1	1	1	1	1	1
2	(1, 4-6); (2, 4-6)	0	0	1	1	1	1	1
3	(3, 1-3); (4, 1-3)	0	0	0	1	1	1	1
4	(3, 4-6); (4, 4-6)	0	0	0	0	1	1	1
5	(5, 1-3); (6, 1-3)	0	0	0	0	0	1	1
6	(5, 4-6); (6, 4-6)	0	0	0	0	0	0	1

(a) Design 2: Units transfer within two adjacent steps

	Time						
Unit	1	2	3	4	5	6	7
1	0	1	1	1	1	1	1
2	0	1	1	1	1	1	1
3	0	1	1	1	1	1	1
4	0	0	1	1	1	1	1
5	0	0	1	1	1	1	1
6	0	0	1	1	1	1	1

(b) Contamination in clusters 1–2

	Time						
Unit	1	2	3	4	5	6	7
1	0	0	0	1	1	1	1
2	0	0	0	1	1	1	1
3	0	0	0	1	1	1	1
4	0	0	0	0	1	1	1
5	0	0	0	0	1	1	1
6	0	0	0	0	1	1	1

(c) Contamination in clusters 3–4

	Time						
Unit	1	2	3	4	5	6	7
1	0	0	0	0	0	1	1
2	0	0	0	0	0	1	1
3	0	0	0	0	0	1	1
4	0	0	0	0	0	0	1
5	0	0	0	0	0	0	1
6	0	0	0	0	0	0	1

(d) Contamination in clusters 5–6

Figure 5.1: Contamination under design 2. Units marked with dashed lines are under the risk of contamination.

		Time						
Pattern	(Cluster, Units)	1	2	3	4	5	6	7
1	(1, 1-3); (2, 1-3)	0	1	1	1	1	1	1
2	(3, 1-3); (4, 1-3)	0	0	1	1	1	1	1
3	(5, 1-3); (6, 1-3)	0	0	0	1	1	1	1
4	(1, 4-6); (2, 4-6)	0	0	0	0	1	1	1
5	(3, 4-6); (4, 4-6)	0	0	0	0	0	1	1
6	(5, 4-6); (6, 4-6)	0	0	0	0	0	0	1

(a) Design 3: Units transfer in two nonadjacent steps

	Time						
Unit	1	2	3	4	5	6	7
1	0	1	1	1	1	1	1
2	0	1	1	1	1	1	1
3	0	1	1	1	1	1	1
4	0	0	0	0	1	1	1
5	0	0	0	0	1	1	1
6	0	0	0	0	1	1	1

(b) Contamination in clusters 1–2

	Time						
Unit	1	2	3	4	5	6	7
1	0	0	1	1	1	1	1
2	0	0	1	1	1	1	1
3	0	0	1	1	1	1	1
4	0	0	0	0	0	1	1
5	0	0	0	0	0	1	1
6	0	0	0	0	0	1	1

(c) Contamination in clusters 3–4

	Time						
Unit	1	2	3	4	5	6	7
1	0	0	0	1	1	1	1
2	0	0	0	1	1	1	1
3	0	0	0	1	1	1	1
4	0	0	0	0	0	0	1
5	0	0	0	0	0	0	1
6	0	0	0	0	0	0	1

(d) Contamination in clusters 5–6

Figure 5.2: Contamination under design 3. Units marked with dashed lines are under the risk of contamination.

		Time						
Pattern	(Cluster, Units)	1	2	3	4	5	6	7
1	(1-6, 1)	0	1	1	1	1	1	1
2	(1-6, 2)	0	0	1	1	1	1	1
3	(1-6, 3)	0	0	0	1	1	1	1
4	(1-6, 4)	0	0	0	0	1	1	1
5	(1-6, 5)	0	0	0	0	0	1	1
6	(1-6, 6)	0	0	0	0	0	0	1

(a) Design 4: Units transfer at all of the steps

	Time						
Unit	1	2	3	4	5	6	7
1	0	1	1	1	1	1	1
2	0	0	1	1	1	1	1
3	0	0	0	1	1	1	1
4	0	0	0	0	1	1	1
5	0	0	0	0	0	1	1
6	0	0	0	0	0	0	1

(b) Contamination in each cluster

Figure 5.3: Contamination under design 4. Units marked with dashed lines are under the risk of contamination.

5.1.3 Results

Figures 5.4–5.7 provides boxplots of the ML estimator of the intervention effect θ when the above four models are fit to data generated from the underlying model (5.1). Note that since contamination is not assumed to occur under design 1, results for design 1 are shown only for fitting the correct model and the model with misspecification of the random unit effect. In addition, when contamination rate is 1, there are not enough comparisons between treatment groups under design 4 when the model correctly accounts for contamination. Hence, results cannot be shown for design 4 when contamination rate is 1 (Figures 5.4 and 5.5). Tables 5.1–5.3 display coverage probabilities of 95% confidence intervals for θ given various values of contamination rate and intensity when fitting the above four models under designs 2, 3, and 4, respectively.

Under design 1 with no contamination, we have that the empirical coverage probabilities when fitting the correct model and the model with misspecification of the random unit effect are the same (0.9490).

When a model correctly accounts for contamination but wrongly assumes data to be three level, there is minimal influence to bias and precision of the ML estimator of the intervention effect θ. Given that contamination is correctly accounted for, the ML estimator of θ is unbiased regardless of contamination rate or intensity (Figure 5.5). Also, the precision of $\hat{\theta}$ is quite close to the precision when fitting the underlying model (Figure 5.4). As for the coverage probability of 95% confidence intervals for θ, the second column in Tables 5.1–5.3 shows that nominal coverage is achieved for each considered scenario and each design.

However, when contamination is not accounted for in model fitting, fitting the standard two-level model (1.7) leads to biased estimation of θ. In extreme cases when contamination rate or intensity is close to 1, severe bias can occur (Figure 5.6). As a result, the coverage probability of 95% confidence intervals for θ can be quite low (Tables 5.1–5.3). In worst case scenarios, the coverage probability is 0.1213 under design 3 and 0.0309 under design 4 when both contamination rate and intensity are 1. Severity of contamination is much higher in design 4 than in designs 2 and 3. Thus, even when contamination rate and intensity are low, the 95% nominal coverage probability under design 4 may not be achieved. On the other hand, in mild conditions of contamination (low rate and low intensity), the nominal coverage probability can be achieved under designs 2 and 3. Even when contamination rate is 1, the coverage probability can still be achieved under design 2 at low intensity. In this case, even when all control patients from the same ward are subject to contamination, mild

contamination intensity is allowed and inference for the intervention effect may still be valid.

Similar results are obtained when the standard three-level model is used to fit data generated from the underlying model, wrongly adding the random unit effect and not accounting for contamination. Hence, the impact of omitting contamination on inference about the intervention effect is often large, while the impact of misspecifying two-level data to be three-level is almost negligible in the setup of our problem.

Lastly, we note that given contamination rate 1 and high intensities, design 3 no longer has higher efficiency than design 2 when contamination is correctly accounted for (Figures 5.4 and 5.5) due to not enough comparisons between the two treatment groups under design 3. Situations are worsened under design 4, since no estimation of θ was produced in this case.

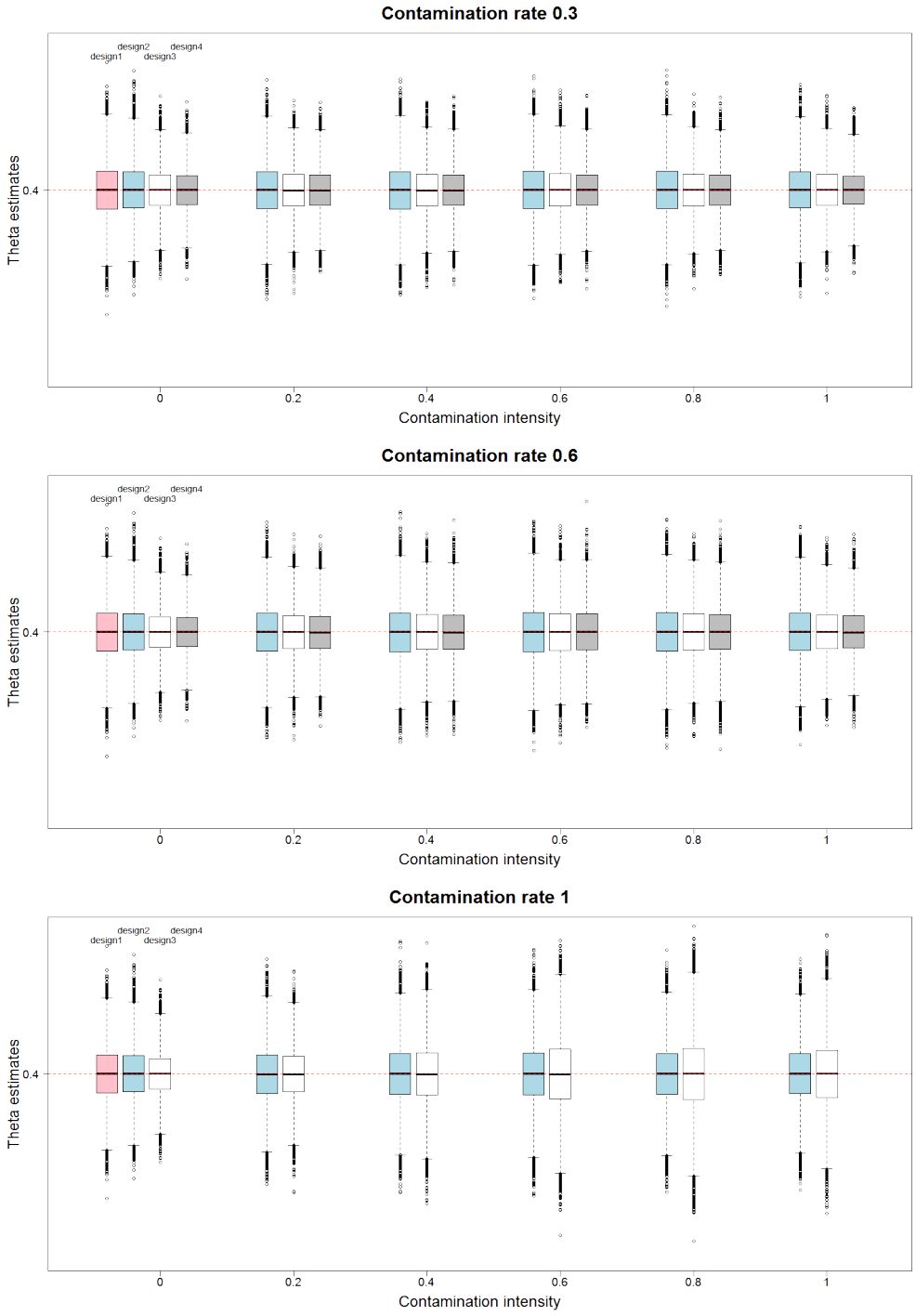


Figure 5.4: Boxplots of $\hat{\theta}$ when fitting the correct model. In the first two panels, results are shown for designs 2, 3 and 4 at all contamination intensities, and for design 1 only when intensity is 0. In the last panel, results are shown for designs 2 and 3 at all contamination intensities and for design 1 only when intensity is 0. No results are shown for design 4 in this case. This figure appears in color in the electronic version of this dissertation.

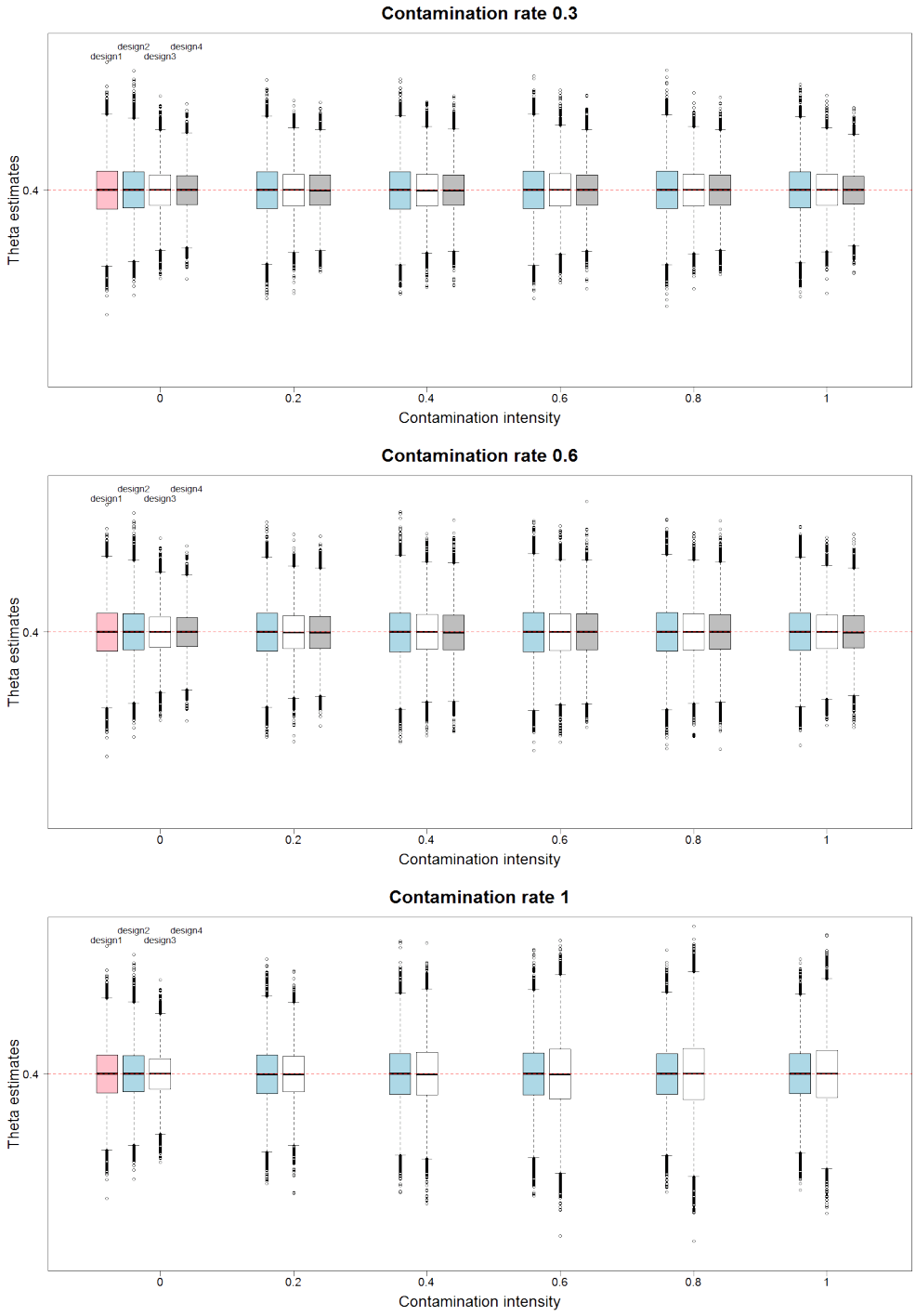


Figure 5.5: Boxplots of $\hat{\theta}$ when fitting model misspecifying the random unit effects. In the first two panels, results are shown for designs 2, 3 and 4 at all contamination intensities, and for design 1 only when intensity is 0. In the last panel, results are shown for designs 2 and 3 at all contamination intensities and for design 1 only when intensity is 0. No results are shown for design 4 in this case. This figure appears in color in the electronic version of this dissertation.

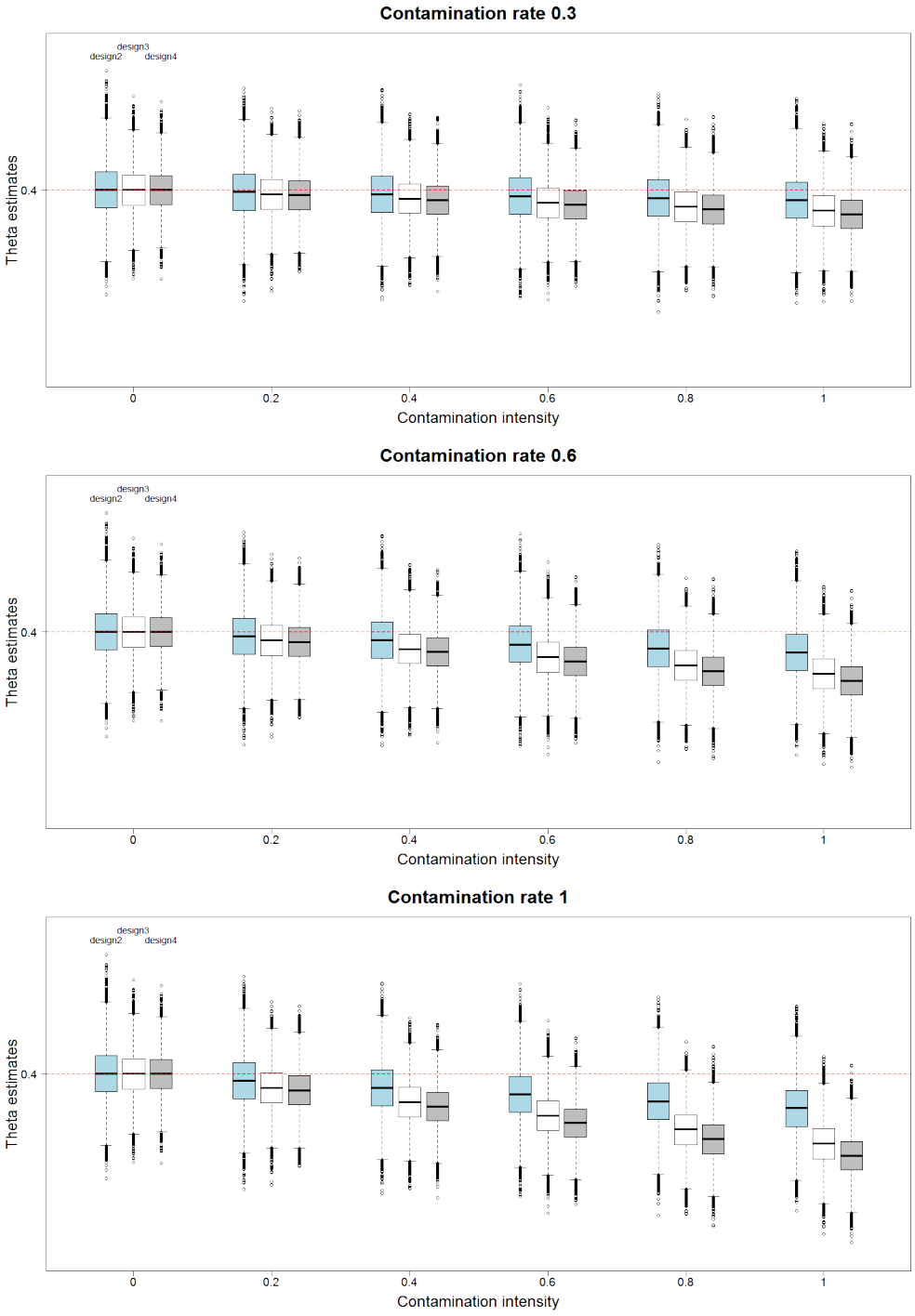


Figure 5.6: Boxplots of $\hat{\theta}$ when fitting model not accounting for contamination. Results are shown for designs 2, 3 and 4 in all panels. No results are shown for design 1. This figure appears in color in the electronic version of this dissertation.

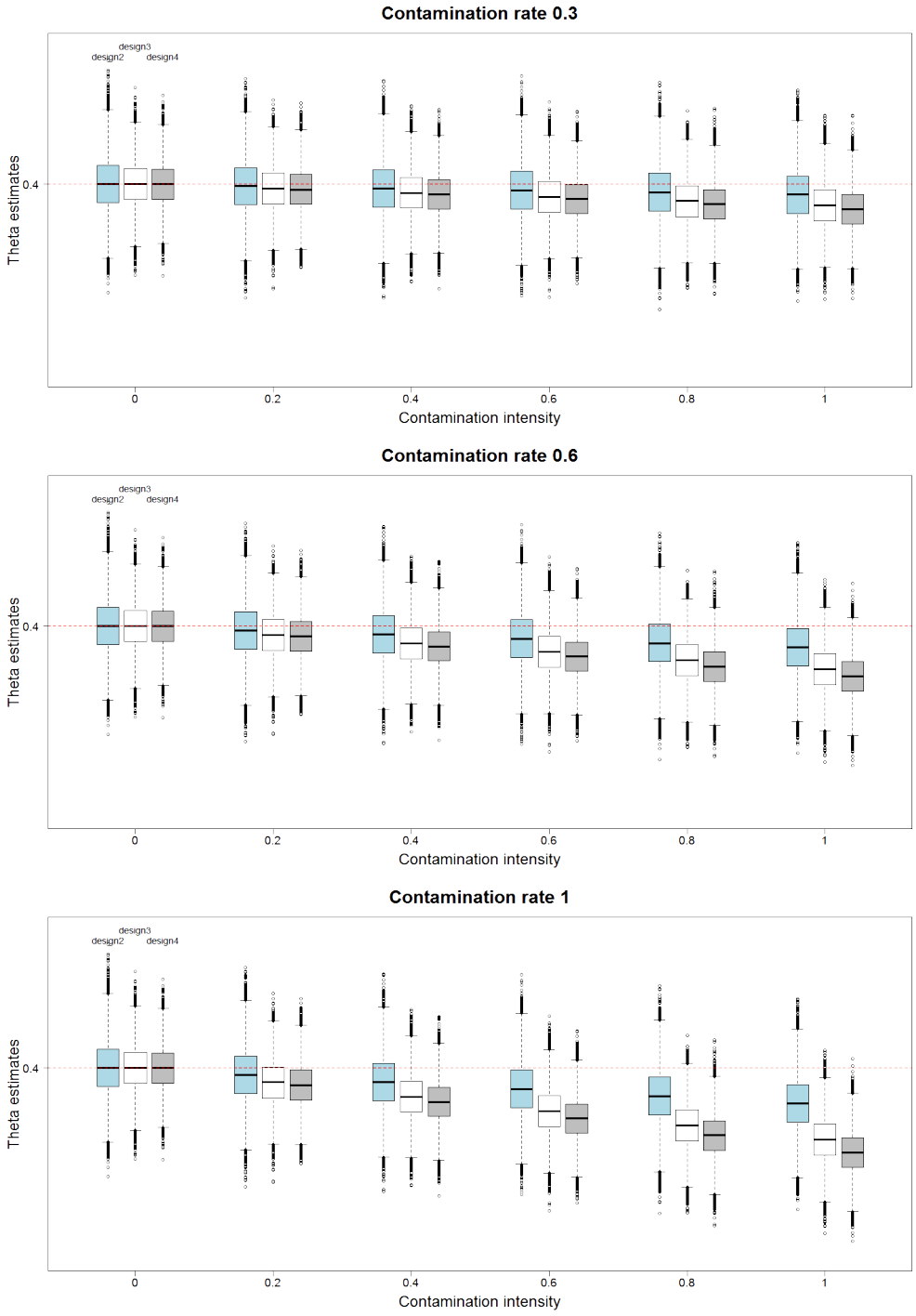


Figure 5.7: Boxplots of $\hat{\theta}$ when fitting the standard three-level model. Results are shown for designs 2, 3 and 4 in all panels. No results are shown for design 1. This figure appears in color in the electronic version of this dissertation.

Table 5.1: Coverage probabilities under design 2

Rate	Intensity	Correct model	Random effects misspecified	Contamination not considered	Standard three-level model
0	0	0.9495	0.9497	–	–
	0.2	0.9522	0.9522	0.9507	0.9510
	0.4	0.9497	0.9501	0.9454	0.9456
0.3	0.6	0.9483	0.9487	0.9416	0.9417
	0.8	0.9515	0.9515	0.9405	0.9405
	1	0.9507	0.9507	0.9308	0.9315
	0.2	0.9513	0.9517	0.9503	0.9506
	0.4	0.9496	0.9495	0.9368	0.9375
0.6	0.6	0.9483	0.9486	0.9236	0.9246
	0.8	0.9509	0.9511	0.9064	0.9074
	1	0.9470	0.9470	0.8786	0.8789
	0.2	0.9511	0.9510	0.9450	0.9455
	0.4	0.9498	0.9498	0.9163	0.9171
1	0.6	0.9476	0.9477	0.8780	0.8787
	0.8	0.9528	0.9526	0.8268	0.8283
	1	0.9488	0.9488	0.7546	0.7567

Table 5.2: Coverage probabilities under design 3

Rate	Intensity	Correct model	Random effects misspecified	Contamination not considered	Standard three-level model
0	0	0.9505	0.9516	–	–
	0.2	0.9515	0.9529	0.9459	0.9482
	0.4	0.9515	0.9528	0.9344	0.9363
0.3	0.6	0.9521	0.9537	0.9150	0.9185
	0.8	0.9494	0.9505	0.8798	0.8825
	1	0.9480	0.9487	0.8448	0.8491
	0.2	0.9512	0.9529	0.9328	0.9349
	0.4	0.9526	0.9536	0.8850	0.8880
0.6	0.6	0.9512	0.9520	0.7956	0.8005
	0.8	0.9504	0.9515	0.6771	0.6851
	1	0.9500	0.9500	0.5315	0.5414
	0.2	0.9510	0.9528	0.9028	0.9059
	0.4	0.9507	0.9516	0.7614	0.7682
1	0.6	0.9510	0.9515	0.5336	0.5428
	0.8	0.9498	0.9505	0.2904	0.3007
	1	0.9508	0.9508	0.1213	0.1277

Table 5.3: Coverage probabilities under design 4

Rate	Intensity	Correct model	Random effects misspecified	Contamination not considered	Standard three-level model
0	0	0.9513	0.9531	–	–
0.3	0.2	0.9494	0.9514	0.9416	0.9444
	0.4	0.9512	0.9528	0.9240	0.9272
	0.6	0.9518	0.9532	0.8927	0.8965
	0.8	0.9480	0.9495	0.8477	0.8528
	1	0.9500	0.9503	0.7880	0.7926
0.6	0.2	0.9496	0.9519	0.9226	0.9250
	0.4	0.9507	0.9521	0.8450	0.8495
	0.6	0.9508	0.9522	0.7123	0.7197
	0.8	0.9506	0.9517	0.5433	0.5521
	1	0.9526	0.9527	0.3603	0.3674
1	0.2	–	–	0.8765	0.8794
	0.4	–	–	0.6595	0.6675
	0.6	–	–	0.3580	0.3663
	0.8	–	–	0.1304	0.1359
	1	–	–	0.0309	0.0326

5.2 Contamination and Model Misspecification by Omitting the Nurse Effect

We consider a scenario where a three-level SW-CRT is conducted to provide improved care to patients. Wards within a hospital are units of randomization. Each nurse may be responsible for caring for over half of patients in a ward and a few patients in other wards. Patients from the same ward may be under the care of different nurses. When a ward of patients is selected to have the crossover to the intervention, all nurses giving care to patients in the ward need to be trained to provide the new care. In this case, the crossed allocation of nurses and wards may inadvertently expose control patients to the intervention. A nurse who cares for patients in intervention

wards may provide care based on the intervention to patients in control wards. Furthermore, investigators may fit the incorrect mixed model that includes the effects of hospitals and wards, omits the nurse effect, and thus omits the corresponding nurse random effect when fitting the model.

In this section, we consider the above scenario and study the influence of contamination induced by crossed allocation of nurses and wards within hospitals while wards are units of randomization. Severity of contamination of control units is discussed under different stepped wedge designs. We use numerical examples to study the impact of omitting contamination on inference about the intervention effect. We also show results of fitting models ignoring the nurse effect while data are clustered at both ward level and nurse level.

5.2.1 The Model

We consider the following model that describes data collected from three-level CRTs with crossed ward and nurse random effects.

$$Y_{ijstk} = \delta_t + \theta Q_{ijstk} + a_i + b_{j(i)} + c_{s(i)} + \epsilon_{ijstk}, \tag{5.2}$$

where $i = 1, \cdots, I$ is the index for independent hospitals, $j = 1, \cdots, J$ is the index for wards nested within the same hospital, $s = 1, \cdots, S$ is the index for nurses within each hospital, $t = 1, \cdots, T$ is the index for time points, and $k = 1, \cdots, K_{js(i)}$ is the index for patients observations collected from the jth ward and the sth nurse within the ith hospital at each time point. Accordingly, Y_{ijstk} denotes the kth observation within the jth ward and sth nurse nested in the ith cluster at time t. δ_t and θ denote the fixed time effect at time t and the intervention effect, respectively. As

with previous chapters, we focus on cross-sectional studies where the observations Y_{ijtk} are collected from $IJTK$ different participants.

The variable Q_{ijstk} is equal to 1 if patient k from ward j and nurse s within hospital i is assigned to the intervention at time t, 0 if the patient is assigned to the control and not contaminated at time t, and a value between 0 and 1 if the patient is assigned to the control but subject to contamination at time t. A value of Q_{ijstk} between 0 and 1 represents the proportion of intervention received by the contaminated control participant. A large value of Q_{ijstk} indicates high contamination intensity.

The hospital random effects $a_i, i = 1, \cdots, I$ are independently drawn from $N(0, \sigma_a^2)$. Within the same hospital i, the ward random effects $b_{j(i)}, j = 1, \cdots, J$ are independently drawn from $N(0, \sigma_b^2)$, and the nurse random effect $c_{s(i)}, s = 1, \cdots, S$ are independently drawn from $N(0, \sigma_c^2)$. The random errors ϵ_{ijtk} are independently drawn from $N(0, \sigma_e^2)$. Lastly, the random effects $a_i, b_{j(i)}, c_{s(i)}$ and the error ϵ_{ijtk} are mutually independent. Hence, we have

- For each observation, the variance is $\text{Var}(Y_{ijstk}) = \sigma_a^2 + \sigma_b^2 + \sigma_c^2 + \sigma_e^2$.
- For two observations from the same ward j and same nurse s within hospital i, $\text{Cov}(Y_{ijstk}, Y_{ijstk'}) = \text{Cov}(Y_{ijstk}, Y_{ijst'k'}) = \sigma_a^2 + \sigma_b^2 + \sigma_c^2, k \neq k', t \neq t'$.
- For two observations from the same ward j but two different nurses s and s' within hospital i, $\text{Cov}(Y_{ijstk}, Y_{ijs'tk'}) = \text{Cov}(Y_{ijstk}, Y_{ijs't'k'}) = \sigma_a^2 + \sigma_b^2, s \neq s', k \neq k', t \neq t'$.
- For two observations from different wards j and j' but same nurse s within hospital i, $\text{Cov}(Y_{ijstk}, Y_{ij'stk'}) = \text{Cov}(Y_{ijstk}, Y_{ij'st'k'}) = \sigma_a^2 + \sigma_c^2, j \neq j', k \neq k', t \neq t'$.

- For two observations not sharing the same ward or nurse within hospital i, $\text{Cov}(Y_{ijstk}, Y_{ij's'tk'}) = \text{Cov}(Y_{ijstk}, Y_{ij's't'k'}) = \sigma_a^2, j \neq j', s \neq s', k \neq k', t \neq t'$.

We define three important concepts based on model (5.2). First, the intra-cluster correlation (ICC) among patient observations within the same ward and same nurse within a hospital $\rho \coloneqq \frac{\sigma_a^2+\sigma_b^2+\sigma_c^2}{\sigma_a^2+\sigma_b^2+\sigma_c^2+\sigma_e^2}$. Second, the ratio of the correlation between observations not sharing wards or nurses within the same hospital to the ICC, $\eta_a \coloneqq \frac{\sigma_a^2}{\sigma_a^2+\sigma_b^2+\sigma_c^2}$; it essentially describes the relative contribution of hospitals to the ICC. Third, the relative contribution of wards within a hospital to the ICC, $\eta_b \coloneqq \frac{\sigma_b^2}{\sigma_a^2+\sigma_b^2+\sigma_c^2}$. We also have $1 - \eta_a - \eta_b = \frac{\sigma_c^2}{\sigma_a^2+\sigma_b^2+\sigma_c^2}$ describing the contribution of nurses within a hospital to the ICC.

We define the variance of Y_{ijtsk} to be $\sigma_y^2 \coloneqq Var(Y_{ijstk}) = \sigma_a^2 + \sigma_b^2 + \sigma_c^2 + \sigma_e^2$. In this case, the variance components $\sigma_a^2, \sigma_b^2, \sigma_c^2$ and σ_e^2 can be uniquely determined by the total variation σ_y^2 in the data and the three important quantities ρ, η_a and η_b.

5.2.2 Problem Setup and Contamination under Each Design

We considered 18 participating hospitals, 6 wards selected within each hospital, and 20 patients recruited in each ward at each phase. Each study was assumed to last for 7 time periods, and patients data were generated during each time period. For each hospital, 6 nurses have varying responsibilities for different wards. We designed a crossed allocation of nurses to wards such that each nurse is responsible for half of patients within a specific ward and fewer patients from other wards. Figure 5.8 shows a possible allocation following our idea. Each of the 6 nurses cares for 10 patients from a certain ward and 2 patients from each of the other 5 wards. With such allocation of nurses to wards, contamination may happen to patients in the control ward

cared for by nurses who also provide intervention care to patients in the intervention ward. This situation yields essentially the same contamination mechanism as in the previous section (Figures 5.1–5.3). Previously, we assumed that all contaminated control patients within the same ward receive the same contamination intensity. In this section, contaminated control patients within the same ward may or may not receive the same contamination intensity depending on which nurse is providing care.

Under designs 2 and 3, wards 4–6 from each hospital have the risk of contamination when wards 1–3 have the crossover to the intervention (Figures 5.1 and 5.2). The intervention is delivered by nurses who have received the new training program. In our problem setup, all six nurses from each hospital receive the new training and deliver intervention care to patients in wards 1–3. Among them, nurses 1–3 have the most intensive interaction with patients within the first three wards, since each of them care for 10 patients in each of the wards. On the other hand, nurses 4–6 have less interaction with patients from wards 1–3. Hence, we assumed that control patients cared for by nurses more intensively interacting with intervention patients are under more severe contamination, and those cared by nurses less intensively interacting with intervention patients are under less severe contamination. Figure 5.9 (a) provides graphic illustration of contamination of control patients from wards 4–6. Patients in wards 4–6 cared for by nurses 1–3 have severe contamination, while patients in wards 4–6 cared for by nurses 4–6 have mild contamination.

Under design 4, contamination intensity of patients in control wards may be different across different times. For example, during time period 2 when ward 1 crosses over to the intervention, patients cared for by nurse 1 in the other wards may have intense contamination, while patients cared for by nurses 2–6 in these wards have

less intense contamination (Figure 5.9 (b)). During time period 3, both ward 1 and ward 2 are in the intervention arm. In this case, patients in wards 3–6 cared for by nurses 1 and 2 have more intense contamination, while patients cared for by nurses 1 and 2 in these wards have less intense contamination (Figure 5.9 (c)). Following this idea, contamination intensity of patients in control wards at time 4, 5, and 6 can also be obtained. Throughout this section, we assumed the contamination rate to be 1, meaning that all control patients are contaminated if they share the same nurse as patients in other wards receiving the intervention.

Nurse \ Ward	1	2	3	4	5	6
1	**10 patients**	2 patients	2 patients	2 patients	2 patients	2 patients
2	2 patients	**10 patients**	2 patients	2 patients	2 patients	2 patients
3	2 patients	2 patients	**10 patients**	2 patients	2 patients	2 patients
4	2 patients	2 patients	2 patients	**10 patients**	2 patients	2 patients
5	2 patients	2 patients	2 patients	2 patients	**10 patients**	2 patients
6	2 patients	2 patients	2 patients	2 patients	2 patients	**10 patients**

Figure 5.8: Scheme of nurse responsibilities across wards within each hospital

We let the total variation of the outcome to be $\sigma_y = 5$. With the underlying model (5.2), we set ρ to be 0.05, η_a to be 0.4, and η_b to be 0.2. The underlying intervention effect θ was determined as 0.4 in order to achieve 80% power of detecting the intervention effect when fitting the correct model (5.2) with the above setup under design 1. For ease of analysis, we assumed no time effects and thus $\delta_1 = \cdots = \delta_T = 50$. To generate data with contamination, we assumed the high intensity to be $c = 0.5$ and the low intensity to be $c = 0.1$. This gives us the ratio of 5 to 1 based on each nurse's high and low contributions to wards. Hence, contaminated control patients may experience an intervention effect as high as 50% and as low as 10%.

Nurse \ Ward	1	2	3	4	5	6
1	**10 patients**	2 patients	2 patients	2 patients	2 patients	2 patients
2	2 patients	**10 patients**	2 patients	2 patients	2 patients	2 patients
3	2 patients	2 patients	**10 patients**	2 patients	2 patients	2 patients
4	2 patients	2 patients	2 patients	**10 patients**	2 patients	2 patients
5	2 patients	2 patients	2 patients	2 patients	**10 patients**	2 patients
6	2 patients	2 patients	2 patients	2 patients	2 patients	**10 patients**

(a) Contamination under designs 2 and 3 at time points when contamination is possible.

Nurse \ Ward	1	2	3	4	5	6
1	**10 patients**	2 patients	2 patients	2 patients	2 patients	2 patients
2	2 patients	**10 patients**	2 patients	2 patients	2 patients	2 patients
3	2 patients	2 patients	**10 patients**	2 patients	2 patients	2 patients
4	2 patients	2 patients	2 patients	**10 patients**	2 patients	2 patients
5	2 patients	2 patients	2 patients	2 patients	**10 patients**	2 patients
6	2 patients	2 patients	2 patients	2 patients	2 patients	**10 patients**

(b) Contamination under design 4 at time point 2.

Nurse \ Ward	1	2	3	4	5	6
1	**10 patients**	2 patients	2 patients	2 patients	2 patients	2 patients
2	2 patients	**10 patients**	2 patients	2 patients	2 patients	2 patients
3	2 patients	2 patients	**10 patients**	2 patients	2 patients	2 patients
4	2 patients	2 patients	2 patients	**10 patients**	2 patients	2 patients
5	2 patients	2 patients	2 patients	2 patients	**10 patients**	2 patients
6	2 patients	2 patients	2 patients	2 patients	2 patients	**10 patients**

(c) Contamination under design 4 at time point 3.

Figure 5.9: Contamination under designs 2–4. Cells filled with dark gray means corresponding patients have more intense contamination. Cells filled with light gray means corresponding patients have less intense contamination.

Under each of designs 1–4, we generated $10,000$ datasets from the underlying model (5.2). In this case, the Monte Carlo error allowed in estimating 95% confidence interval coverage probabilities is $\sqrt{0.95(1-0.95)/10000}$, approximately 0.0022.

As with the previous section, our analysis includes four steps. First, we fit the data-generating model (5.2) as a reference. Second, we fit a model that correctly incorporates contamination rate and intensity of control subjects, but wrongly ignores

the nurse random effect $c_{s(i)}$, using

$$Y_{ijstk} = \delta_t + \theta Q_{ijstk} + a_i + b_{j(i)} + \epsilon_{ijstk},$$

where all notations follow from the true model. Third, we fit a model that correctly specifies the random effects, but fails to acknowledge contamination happened to control subjects, using

$$Y_{ijstk} = \delta_t + Z_{ijt}\theta + a_i + b_{j(i)} + c_{s(i)} + \epsilon_{ijstk},$$

where the treatment indicator Z_{ijt} follows from the original study design. Lastly, we fit a model both ignoring the nurse random effect and failing to acknowledge contamination, using

$$Y_{ijstk} = \delta_t + Z_{ijt}\theta + a_i + b_{j(i)} + \epsilon_{ijstk},$$

where Z_{ijt} follows from the original study design and all notations follow from the underlying model (5.2).

5.2.3 Results

When fitting the true data-generating model, unbiased estimation of θ and the 95% coverage probability is achieved under all designs as expected (Tables 5.4 and 5.5). On the other hand, when contamination is not accounted for, bias and low coverage probability of the 95% confidence intervals for θ occur under designs 2–4. Although the estimation of θ is biased under design 2, the coverage probability is only slightly under 0.95. On the contrary, designs 3 and 4 yield more biased estimation and lower coverage probability for θ even though they are more efficient in this case.

When contamination intensities are correctly specified, the order of efficiency among the four designs is design 1 > design 2 > design 3 > design 4 (Figure 5.4

and Figure 5.5). As control patients are contaminated, their responses may become similar to responses of intervention patients. Accordingly, under designs 2–4, comparisons of responses across uncontaminated control patients, contaminated control patients, and intervention patients may no longer be sharp. In worse case scenarios when both contamination rate and intensity are high, little information can be used for estimating the intervention effect under these three designs. On the contrary, design 1 provides sharp contrast in patients responses between the two trial arms, since all comparisons are "0–1" (between uncontaminated control patients and intervention patients).

Similar results are obtained from fitting the model that correctly accounts for contamination but omits the nurse random effect. The estimation of θ is unbiased under each of the four designs, with lowest precision under design 4 and highest precision under design 1. The impact of random effect misspecification is minimal, especially for designs 1–3. Under design 4, the coverage probability is slightly below 0.95 due to high standard error of the ML estimator $\hat{\theta}$ (0.162).

Comparing the last two columns in Tables 5.4 and 5.5, we again find that ignoring the nurse random effect has minimal impact on inference of θ. Only slight increase in the standard error of $\hat{\theta}$ is noticed under design 4, which in turn affects the coverage probability for θ.

Table 5.4: Empirical means and standard errors of $\hat{\theta}$

	Correct model	Random effects misspecified	Contamination not considered	Standard three-level model
Design 1	0.4012(0.137)	0.4012(0.137)	–	–
Design 2	0.4047(0.145)	0.4046(0.146)	0.3691(0.137)	0.3697(0.138)
Design 3	0.4010(0.147)	0.4008(0.154)	0.3254(0.128)	0.3299(0.130)
Design 4	0.3974(0.152)	0.3975(0.162)	0.3045(0.125)	0.3097(0.127)

Table 5.5: Coverage probabilities for θ

	Correct model	Random effects misspecified	Contamination not considered	Standard three-level model
Design 1	0.9531	0.9544	–	–
Design 2	0.9478	0.9463	0.9424	0.9441
Design 3	0.9490	0.9392	0.9034	0.9144
Design 4	0.9470	0.9316	0.8769	0.8906

5.3 Discussion

In this chapter, we have investigated two interesting scenarios that could potentially lead to contamination under designs 2–4. We have varied the severity of contamination in terms of rate and intensity. Section 5.1 comprehensively investigated consequences of not correctly accounting for contamination for various levels of contamination severity. In conditions of low contamination rate and intensity, designs 2 and 3 can still retain the 95% nominal coverage probability even when the model being fitted does not account for contamination. Our finding is consistent with the conclusion in Schochet (2008) stating designs that randomize within clusters are appropriate when spillover effects are expected to be small. On the other hand, design 4 can not maintain the nominal coverage probability when contamination was present.

Section 5.2 considered crossed allocation of nurses to wards leading to potential contamination of patients in control wards. All control patients cared for by nurses who also give care to intervention patients were assumed to be contaminated (contamination rate equal to 1). For fixed nurse contributions to each ward, contamination rates, and intensities, only design 2 almost retains the 95% coverage probability when contamination is not accounted for. Under-coverage happens in designs 3 and 4 in this case.

Our results also show that in the presence of contamination, study designs randomizing wards within hospitals are still promising. If investigators believe that contamination between wards within the same hospital is mild, study designs 2 and 3 can be carried out and provide intervention effect estimates with high efficiency and 95% coverage probability of confidence intervals. When contamination is mild to moderate, design 2 may still be chosen for retaining the 95% coverage probability.

We have also studied the consequence of misspecifying the random effects in models (5.1) and (5.2). Results show that neither wrongly adding the random unit effect in model (5.1) nor omitting the nurse random effect in model (5.2) has noticeable impact on inference of θ. Our results agree with the conclusions in McCulloch and Neuhaus (2011).

Our analyses can be improved in several aspects. First, we may also examine the power of detecting the intervention effect when contamination causes dilution of the effect size (Torgerson, 2001; Rhoads, 2011). As discussed before, contamination severity and design efficiency varies under different designs. Design 1 has lowest risk of contamination and lowest efficiency, and design 4 has highest risk of contamination and highest efficiency. When there is risk of contamination, power of detecting the

intervention effect depends on both severity of contamination and design efficiency. Second, it would be interesting to consider other possible allocation of nurses' contributions to wards. In Section 5.2, we assumed that each of the six nurses within a hospital contributes 50% to one ward and 10% to each of the other five wards. Alternatively, a nurse may devote 80% of the total contribution to a single ward, and 20% to another ward. In this case, the risk of contamination may be lowered when wards are units of randomization. Last, it may also be interesting to study the consequences of accounting for contamination but in a wrong way. For example, robustness of inference about the intervention effect can be studied when contamination rate is inaccurately estimated.

Chapter 6: Summary and Future Work

6.1 Summary

In this dissertation, we proposed novel stepped wedge cluster randomized trials for three-level data. The proposed designs differ in timing of allocating units within the same cluster to different treatments. We evaluated the efficiency of each design under a variety of underlying models generating three-level data. Impacts of misspecifying random unit effects and omitting contamination on inference about the intervention effect were also evaluated via simulation studies.

We derived the closed-form expression for power of testing the intervention effect under the standard three-level model (2.1) incorporating time effects. Our power formula is flexible and can be applied to arbitrary types of three-level CRTs. A key component in obtaining the power formula is to derive the closed-form expression for the variance of the GLS estimator $\hat{\theta}$ for the intervention effect. This expression also enables researchers to analyze the precision in estimating the intervention effect for both routine and non-routine three-level CRTs without the need for simulation.

Our proposed three-level SW-CRTs differ in timing of allocating units within the same cluster to different treatments. In design 1, all units from the same cluster

transfer at a single time point. In designs 2 and 3, units from the same cluster transfer at two adjacent time points and two nonadjacent time points, respectively. In design 4, units from the same cluster transfer at all time points when the one-way crossover happens. Using the derived variance formula under the standard three-level model in Chapter 2, we computed and compared the efficiency of the four designs given varying ICC ρ and the ratio of between-unit to within-unit correlations η. To explain the patterns in variance of the treatment effect under different scenarios, we identified three types of comparisons involved in estimating the treatment effect under the standard three-level model. Large correlations between observations within the same unit/cluster improve the precision of intervention effect estimates via within-unit/within-cluster comparisons (types 1 and 2), and decrease the precision of intervention effect estimates via between-cluster comparisons (type 3). For fixed values of ρ and η, the order in efficiency of the four designs is design 4 $>$ design 3 $>$ design 2 $>$ design 1, as design 4 has the most between-unit within-time comparisons (included in type 2 comparisons) and design 1 has no such comparisons.

We considered alternative underlying models to generate data from three-level SW-CRTs. One extended model includes varying random unit effects across time points. Motivated by the example of nurses within hospitals, we believed that the effect of a nurse on a patients outcome variable at one time point may not be the same as the effect of the same nurse at a different time point. We chose Toeplitz covariance pattern for the correlated random unit effects $b_{ijt}, t = 1, \cdots, T$ to generate data from this model. The order in efficiency of the four designs remained the same as that under the standard model. However, patterns in variance of the intervention effect estimator for varying correlations can be quite different depending on how close the

extended model is to the standard model. In addition, we studied the performance of the standard model in this setting. Our numerical studies show that the empirical coverage probability of 95% confidence intervals for θ retains the nominal coverage probability when the wrongly specified model is close to the data-generating model. This finding is true for all designs. On the other hand, when the two models are not close, the nominal coverage probability cannot be retained under any design. However, the estimation of θ remains unbiased in all scenarios considered in this study.

Other ways of misspecifying random effects in three-level models were also considered. These studies include wrongly assuming a random unit effect when none are present, and omitting a random unit effect when it exists. These misspecifications had minimal impact on inference about the intervention effect.

Finally, we addressed the problem of contamination, which could occur in our proposed stepped wedge designs because units from the same cluster crossover at different times. We described severity of contamination using rate (proportion of participants receiving the control that are exposed to the intervention) and intensity (proportion of the intervention effect received by contaminated control participants). In general, design 1 may not induce any contamination between units within the same cluster, and designs 2, 3, and 4 may have minimal, moderate and severe contamination, respectively. Our simulation studies showed that in the presence of mild to moderate contamination, designs 2 and 3 can still be carried out to provide flexibility to study planners while retaining the nominal coverage probability for the treatment effect. On the other hand, design 4 is not an appropriate choice even when contamination is

mild, as it has low empirical performance in maintaining nominal coverage probability for the treatment effect.

6.2 Future Directions

We briefly introduce several interesting future directions beyond this dissertation. First, we consider incomplete three-level SW-CRTs that are extensions of our proposed designs. Second, we discuss several modeling alternatives that may also be underlying models generating real-world data from SW-CRTs. Third, we speculate on cohort designs that may also fit into the framework of SW-CRTs.

6.2.1 Incomplete Three-level SW-CRTs

We introduce two incomplete designs to illustrate how our work can be extended to the incomplete SW-CRT. In a complete SW-CRT, the duration of the trial is the same for each participating units and clusters – from the beginning of the trial to the end of the trial. A potential pitfall is that such long duration of trials may place heavy burden on investigators and care givers, which in turn may negatively influence quality of data (Friedman et al., 2010). Alternatively, some authors (Hemming et al., 2015b) have suggested incomplete SW-CRTs where for each cluster or unit, there is at least one time period when data collection is not needed.

Recall that in design 2, units from the same cluster have the one-way crossover to the intervention at two adjacent time points (Figure 3.1 (b)). For two units in the same cluster with different schedules of treatment transition, there are still multiple overlapping time periods when they receive the same treatment. To ease data collection, we may reduce the number of such overlapping time periods while not

reducing the number of within-cluster comparisons. Figure 6.1 shows a possible incomplete stepped wedge design. For each participating unit, data collection lasts for three time periods, including one period when the control is provided and two periods when the intervention is provided. For two units in the same cluster following different schedules (for example, unit 1 and unit 4 in cluster 1), there is one overlapping period when different treatments are provided and one overlapping period when the intervention is provided to both units.

Following the same idea, we may extend design 3 to incomplete stepped wedge designs (Figures 6.2 and 6.3). In Figure 6.2, data collection lasts for four time periods in each participating unit. For two units in the same cluster following different schedules (for example, unit 1 and unit 4 in cluster 1), there are three overlapping periods when different treatments are provided and no overlapping time periods when the same treatment is provided to both units. This design is beneficial when η is large so that the between-unit within-cluster comparison (type 2) is most efficient. On the other hand, in cases when η is small, we may reduce the number of overlapping periods between units in the same cluster receiving different treatments, as the between-unit within-cluster comparison (type 2) may no longer be very efficient. In doing so, we may be able to further reduce the duration of data collection in each unit, placing less burden on care givers and investigators. Figure 6.3 gives an example illustrating the above idea. In this design scheme, data collection only lasts for three time periods in each participating unit. For two units in the same cluster following different schedules (for example, unit 1 and unit 4 in cluster 1), there is only one overlapping period when different treatments are provided.

	Time							
(Cluster, Units)	1	2	3	4	5	6	7	8
(1-2, 1-3)	0	1	1	.	.	.	.	.
(1-2, 4-6)	.	0	1	1	.	.	.	.
(3-4, 1-3)	.	.	0	1	1	.	.	.
(3-4, 4-6)	.	.	.	0	1	1	.	.
(5-6, 1-3)	.	.	.	.	0	1	1	.
(5-6, 4-6)	.	.	.	.	.	0	1	1

Figure 6.1: Incomplete designs extended from design 2. In each cell, 0 means receiving the control, and 1 means receiving the intervention.

	Time						
(Cluster, Units)	1	2	3	4	5	6	7
(1-2, 1-3)	0	1	1	1	.	.	.
(3-4, 1-3)	.	0	1	1	1	.	.
(5-6, 1-3)	.	.	0	1	1	1	.
(1-2, 4-6)	.	0	0	0	1	.	.
(3-4, 4-6)	.	.	0	0	0	1	.
(5-6, 4-6)	.	.	.	0	0	0	1

Figure 6.2: Incomplete designs extended from design 3 with each unit lasting for 4 periods. In each cell, 0 means receiving the control, and 1 means receiving the intervention.

	Time						
(Cluster, Units)	1	2	3	4	5	6	7
(1-2, 1-3)	0	1	1	.	.	.	.
(3-4, 1-3)	.	0	1	1	.	.	.
(5-6, 1-3)	.	.	0	1	1	.	.
(1-2, 4-6)	.	.	0	0	1	.	.
(3-4, 4-6)	.	.	.	0	0	1	.
(5-6, 4-6)	.	.	.	.	0	0	1

Figure 6.3: Incomplete designs extended from design 3 with each unit lasting for 3 periods. In each cell, 0 means receiving the control, and 1 means receiving the intervention.

6.2.2 Other directions

There are other data-generating models worth considering, such as treatment by time interactions. In practice, it is very likely that treatment effect is different across time points, especially for stepped wedge trials typically lasting for a long time. Second, there may be heterogeneous treatment effect across clusters (Rhoads, 2011). In this case, we may consider treatment by cluster interactions as random effects. Third, other information such as patient-level covariates may also be used.

Finally, we have focused on cross-sectional studies where each patient is measured once, as is typically true for SW-CRTs. However, cohort studies where each patient is followed for multiple times can also fit in the framework of stepped wedge designs (Bennett et al., 2013; Barker et al., 2016). In this case, data may even be four-level due to repeated measurements on the same individual.

Appendix A: Proofs of Theorem 1 and Theorem 2

A.1 Preliminaries

We first introduce the following lemmas that will be used in our proofs.

Lemma 1. *(Theorem 8.5.11 in Harville, 2012) Let*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

represent a partitioned $n \times n$ matrix whose blocks are $A_{11} \in \mathbb{R}^{n_1 \times n_1}$, $A_{12} \in \mathbb{R}^{n_1 \times n_2}$, $A_{21} \in \mathbb{R}^{n_2 \times n_1}$, and $A_{22} \in \mathbb{R}^{n_2 \times n_2}$. Suppose that A_{11} is nonsingular. Then, A is invertible if and only if the Schur complement $A_{22} - A_{21}A_{11}^{-1}A_{12}$ of A_{11} is invertible, and

$$A^{-1} = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \\ -(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{pmatrix} \tag{A.1}$$

Lemma 2. *(3.2.4 in Petersen and Pedersen, 2012) Let $A \in \mathbb{R}^{n \times n}$, and $b, c \in \mathbb{R}^n$. Assume that A is invertible, and that $\mathbf{1}_n + c^T A^{-1} b \neq 0$. Then, we have*

$$(A + bc^T)^{-1} = A^{-1} - \frac{A^{-1}bc^T A^{-1}}{1 + c^T A^{-1} b}.$$

Lemma 3. *Let $a, b \in \mathbb{R}$. Assume that $a \neq 0$ and $a + nb \neq 0$. Then, the inverse of $aI_n + bJ_n$ is*

$$\frac{1}{a}(I_n - \frac{b}{a + nb} J_n).$$

A.2 Proof of Theorem 1

Proof. Given the partitioned design matrix $X = (X_1, x_2)$, we have

$$X^T V^{-1} X = \begin{pmatrix} X_1^T V^{-1} X_1 & X_1^T V^{-1} x_2 \\ x_2^T V^{-1} X_1 & x_2^T V^{-1} x_2 \end{pmatrix}.$$

Note that $x_2^T V^{-1} x_2$ is a scalar. Define $A_{11} \coloneqq X_1^T V^{-1} X_1, A_{12} \coloneqq X_1^T V^{-1} x_2, A_{21} \coloneqq x_2^T V^{-1} X_1$, and $A_{22} \coloneqq x_2^T V^{-1} x_2$. We directly apply Lemma 1 to compute the $(2,2)$th entry of matrix $(X^T V^{-1} X)^{-1}$ in partitioned form, which yields the expression of $\mathrm{Var}(\hat{\beta}_p)$ provided in Theorem 1. □

A.3 Proof of Theorem 2

Proof. First, we derive the closed form of the covariance matrix V under model (2.2). Since $V = I_I \otimes W$ is block diagonal, it suffices to derive the inverse of W. Recall that

$$\begin{aligned} W &\coloneqq I_J \otimes (\sigma^2 I_T + \sigma_b^2 J_T) + \sigma_a^2 J_{JT} \\ &= I_J \otimes (\sigma^2 I_T + \sigma_b^2 J_T) + (\sigma_a \mathbf{1}_{JT})(\sigma_a \mathbf{1}_{JT})^T. \end{aligned} \tag{A.2}$$

Define $A \coloneqq I_J \otimes (\sigma^2 I_T + \sigma_b^2 J_T)$, $b = c = \sigma_a \mathbf{1}_{JT}$. We can compute the inverse of W using Lemma 2:

$$W^{-1} = A^{-1} - \frac{\sigma_a^2 A^{-1} \mathbf{1}_{JT} \mathbf{1}_{JT}^T A^{-1}}{1 + \sigma_a^2 \mathbf{1}_{JT}^T A^{-1} \mathbf{1}_{JT}}. \tag{A.3}$$

By Lemma 3, we have

$$\begin{aligned} A^{-1} &= I_J \otimes (\sigma^2 I_T + \sigma_b^2 J_T)^{-1} \\ &= I_J \otimes \frac{1}{\sigma^2} (I_T - \frac{\sigma_b^2}{\sigma^2 + T\sigma_b^2} J_T) \\ &= I_J \otimes p(I_T - q J_T), \end{aligned} \tag{A.4}$$

where $p \coloneqq \frac{1}{\sigma^2}, q \coloneqq \frac{\sigma_b^2}{\sigma^2 + T\sigma_b^2}$, as defined in Theorem 2. Hence,

$$\begin{aligned} \frac{\sigma_a^2 A^{-1} \mathbf{1}_{JT} \mathbf{1}_{JT}^T A^{-1}}{1 + \sigma_a^2 \mathbf{1}_{JT}^T A^{-1} \mathbf{1}_{JT}} &= \frac{\sigma_a^2}{(\sigma^2 + T\sigma_b^2 + JT\sigma_a^2)(\sigma^2 + T\sigma_b^2)} J_{JT} \\ &= r J_{JT}, \end{aligned} \tag{A.5}$$

where $r \coloneqq \frac{\sigma_a^2}{(\sigma^2+T\sigma_b^2+JT\sigma_a^2)(\sigma^2+T\sigma_b^2)}$, as defined in Theorem 2.

Substituting (A.4) and (A.5) into (A.3) yields

$$W^{-1} = pI_{JT} - pqI_J \otimes J_T - rJ_{JT}.$$

Thus, the inverse of V is

$$\begin{aligned} V^{-1} &= I_I \otimes (pI_{JT} - pqI_J \otimes J_T - rJ_{JT}) \\ &= pI_{IJT} - pqI_{IJ} \otimes J_T - rI_I \otimes J_{JT}. \end{aligned} \tag{A.6}$$

Second, we derive the closed form of the inverse of $X_1^T V^{-1} X_1$. Recall that $X_1 = \mathbf{1}_{IJ} \otimes I_T$ and the form of V is in (A.6). We have

$$X_1^T V^{-1} X_1 = IJ(pI_T - (pq + rJ)J_T),$$

where p, q and r are defined as in Theorem 2. Applying Lemma 3 again, we have

$$(X_1^T V^{-1} X_1)^{-1} = \frac{1}{IJp}(I_T + \frac{pq + rJ}{p - T(pq + rJ)} J_T). \tag{A.7}$$

Third, we derive the closed form expression for the variance of $\hat{\theta}$ under model (2.2). By Theorem 1, we have

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{z^T V^{-1} z - z^T V^{-1} X_1 (X_1^T V^{-1} X_1)^{-1} X_1^T V^{-1} z}. \tag{A.8}$$

Based on the derived expressions of V^{-1} in (A.6) and $(X_1^T V^{-1} X_1)^{-1}$ in (A.7), it is easy to compute $z^T V^{-1} z$ and $z^T V^{-1} X_1 (X_1^T V^{-1} X_1)^{-1} X_1^T V^{-1} z$. Thus, we have

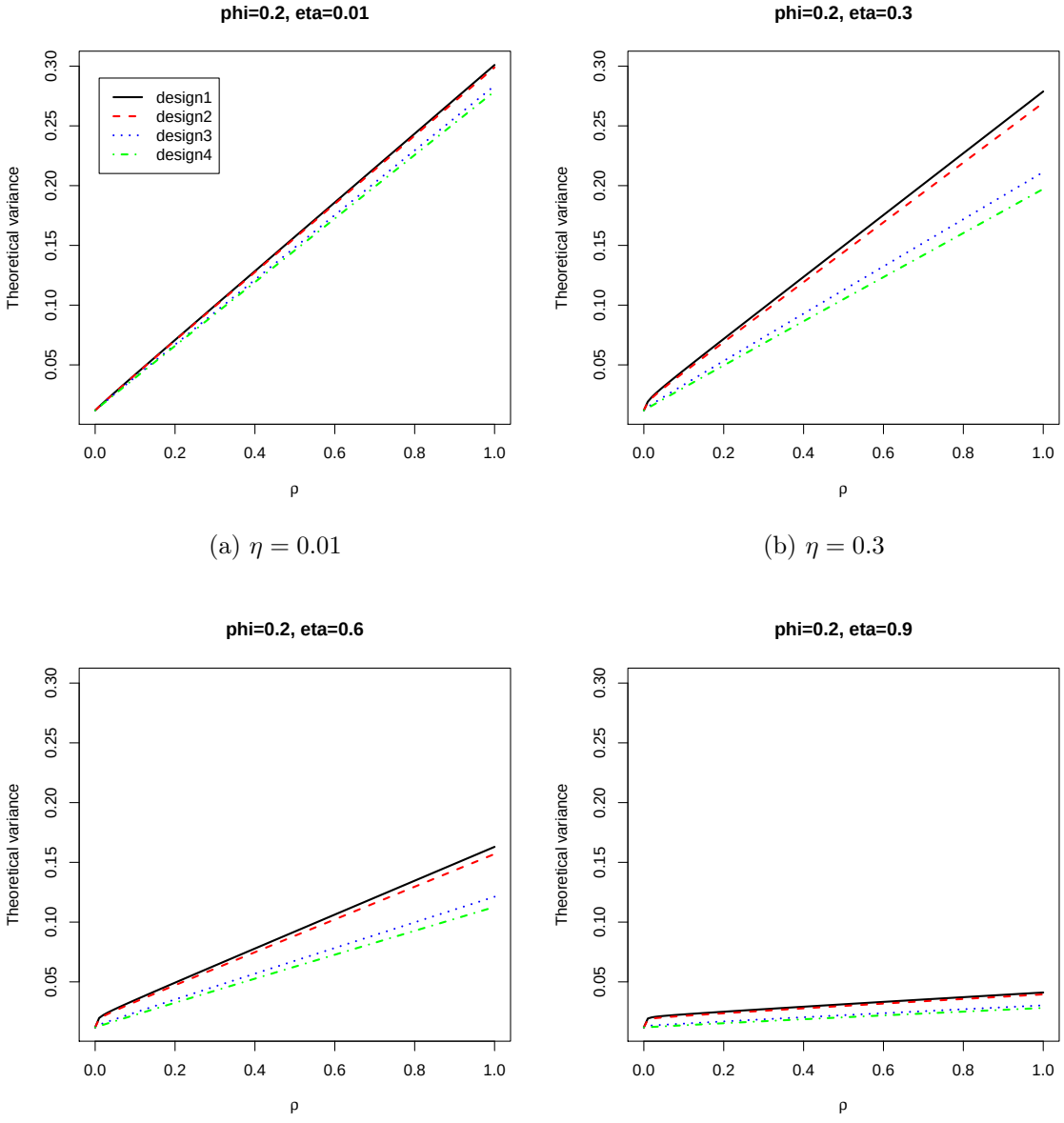
$$z^T V^{-1} z = p \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} Z_{ijt}^2 - pq \sum_{i=1}^{I} \sum_{j=1}^{J} (\sum_{t=1}^{T} Z_{ijt})^2 - r \sum_{i=1}^{I} (\sum_{j=1}^{J} \sum_{t=1}^{T} Z_{ijt})^2. \tag{A.9}$$

and

$$z^T V^{-1} X_1 (X_1^T V^{-1} X_1)^{-1} X_1^T V^{-1} z = \frac{p}{IJ} \sum_{t=1}^{T} (\sum_{i=1}^{I} \sum_{j=1}^{J} Z_{ijt})^2 - \frac{pq + rJ}{IJ} (\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} Z_{ijt})^2 \tag{A.10}$$

We then substitute (A.9) and (A.10) into (A.8). This completes the proof. □

Appendix B: Patterns in Theoretical Variance of Estimator of Intervention Effect When Model (4.1) is Correctly Specified



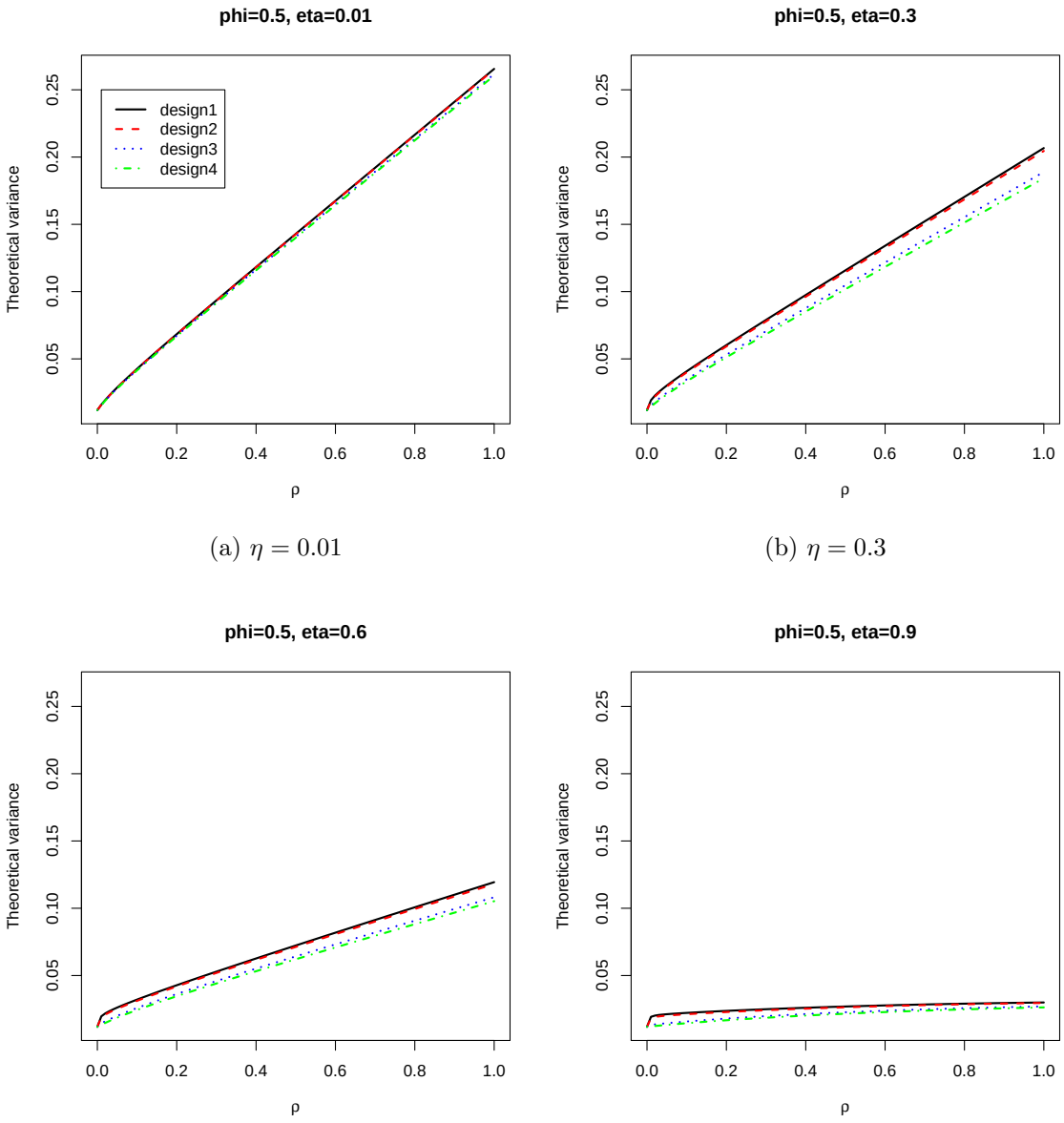
(a) $\eta = 0.01$

(b) $\eta = 0.3$

(c) $\eta = 0.6$

(d) $\eta = 0.9$

Figure B.1: $\text{Var}(\hat{\theta})$ vs. ρ ($\phi_b = 0.2$) when model (4.1) is correctly specified.



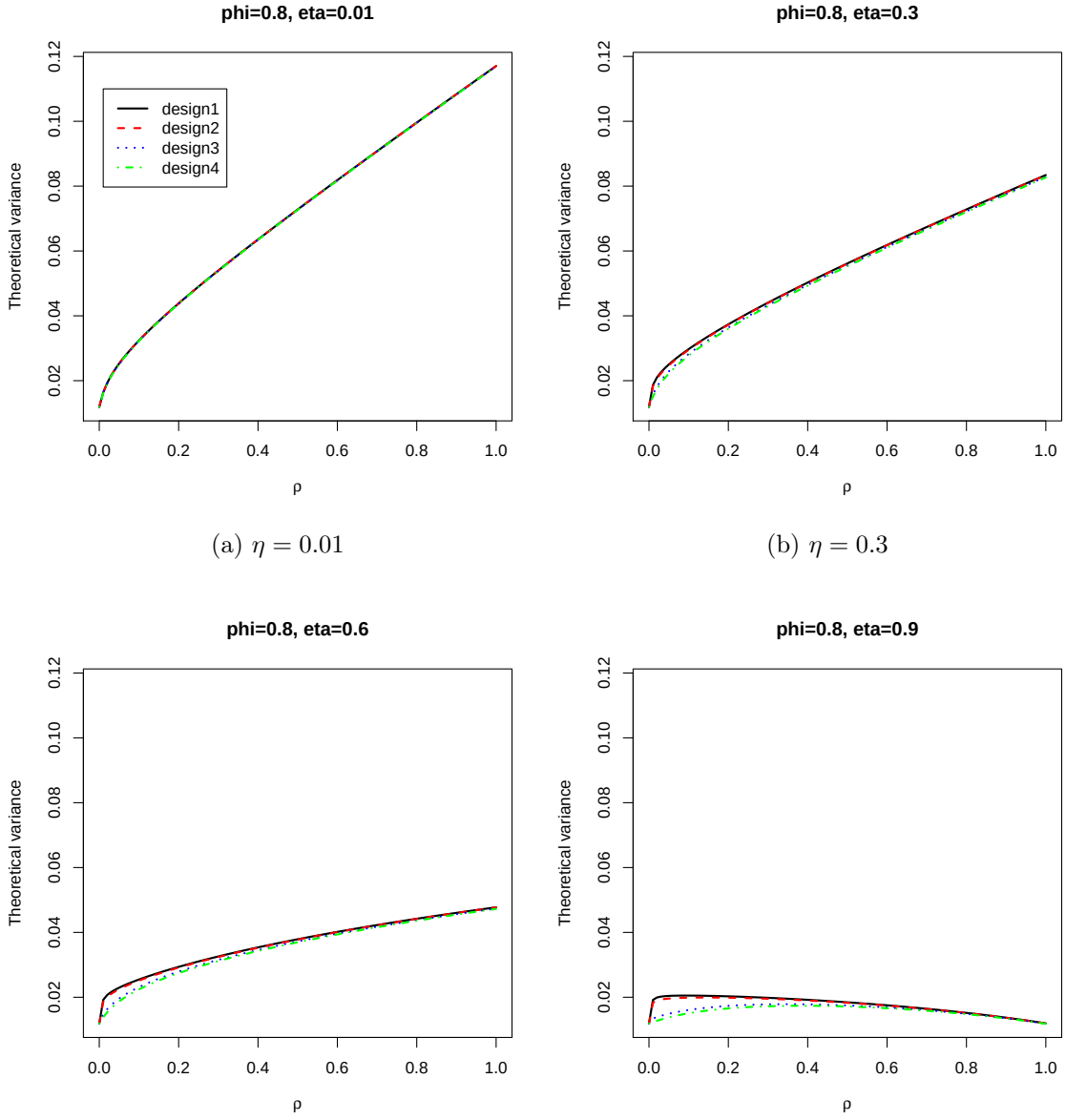
(a) $\eta = 0.01$

(b) $\eta = 0.3$

(c) $\eta = 0.6$

(d) $\eta = 0.9$

Figure B.2: $\text{Var}(\hat{\theta})$ vs. ρ ($\phi_b = 0.5$) when model (4.1) is correctly specified.



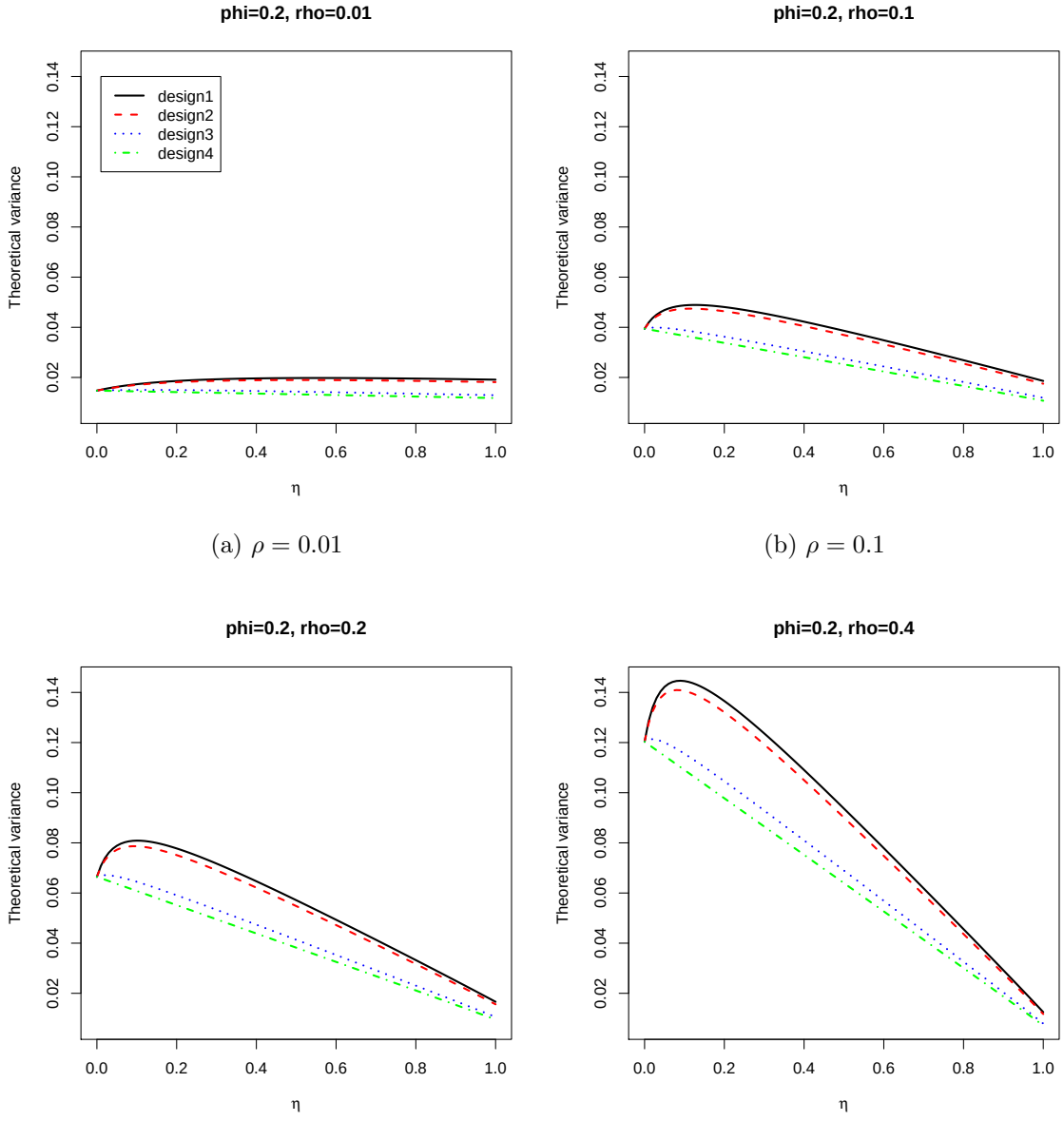
(a) $\eta = 0.01$

(b) $\eta = 0.3$

(c) $\eta = 0.6$

(d) $\eta = 0.9$

Figure B.3: $\text{Var}(\hat{\theta})$ vs. ρ ($\phi_b = 0.8$) when model (4.1) is correctly specified.



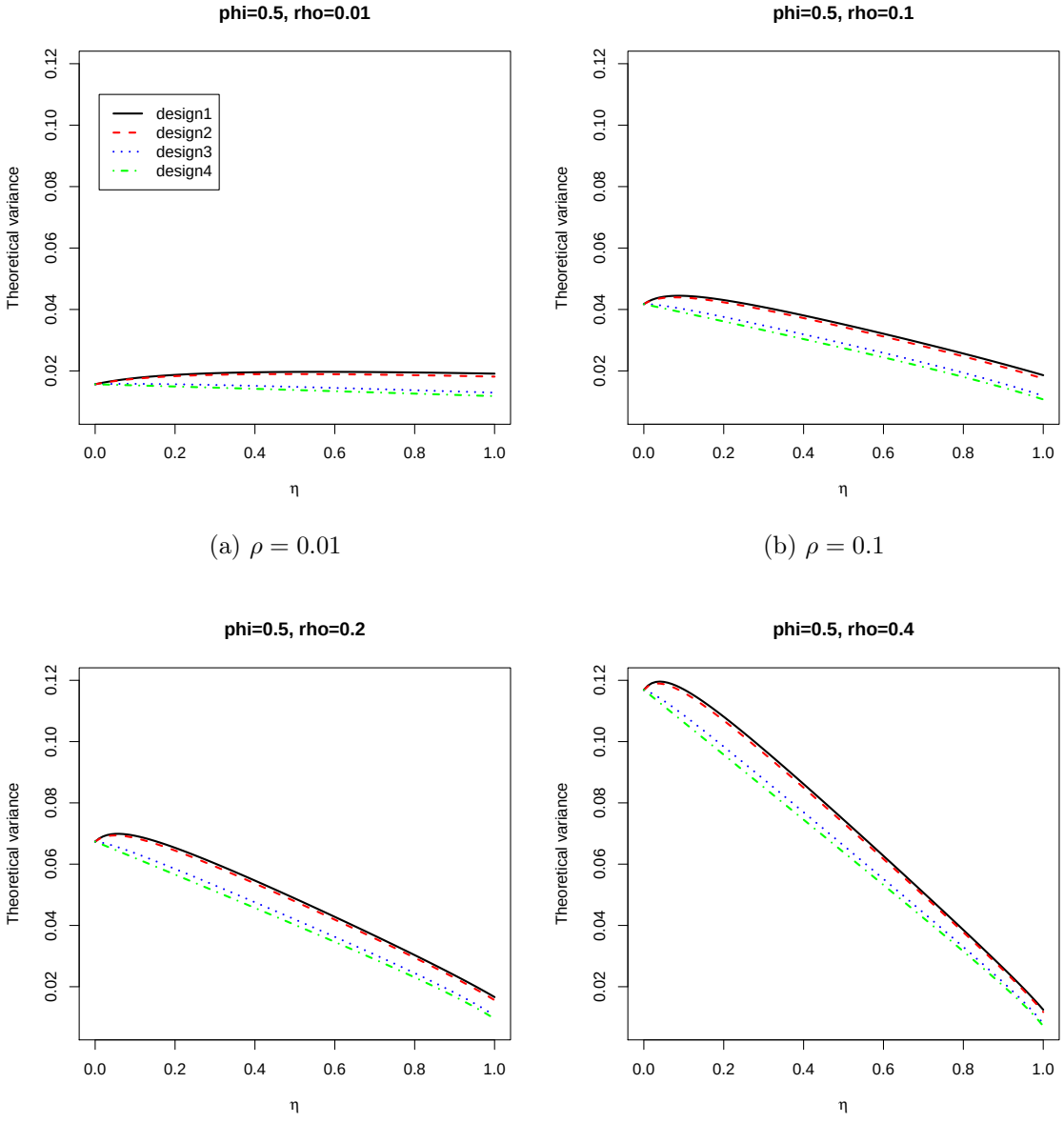
(a) $\rho = 0.01$

(b) $\rho = 0.1$

(c) $\rho = 0.2$

(d) $\rho = 0.4$

Figure B.4: $\text{Var}(\hat{\theta})$ vs. η ($\phi_b = 0.2$) when model (4.1) is correctly specified.



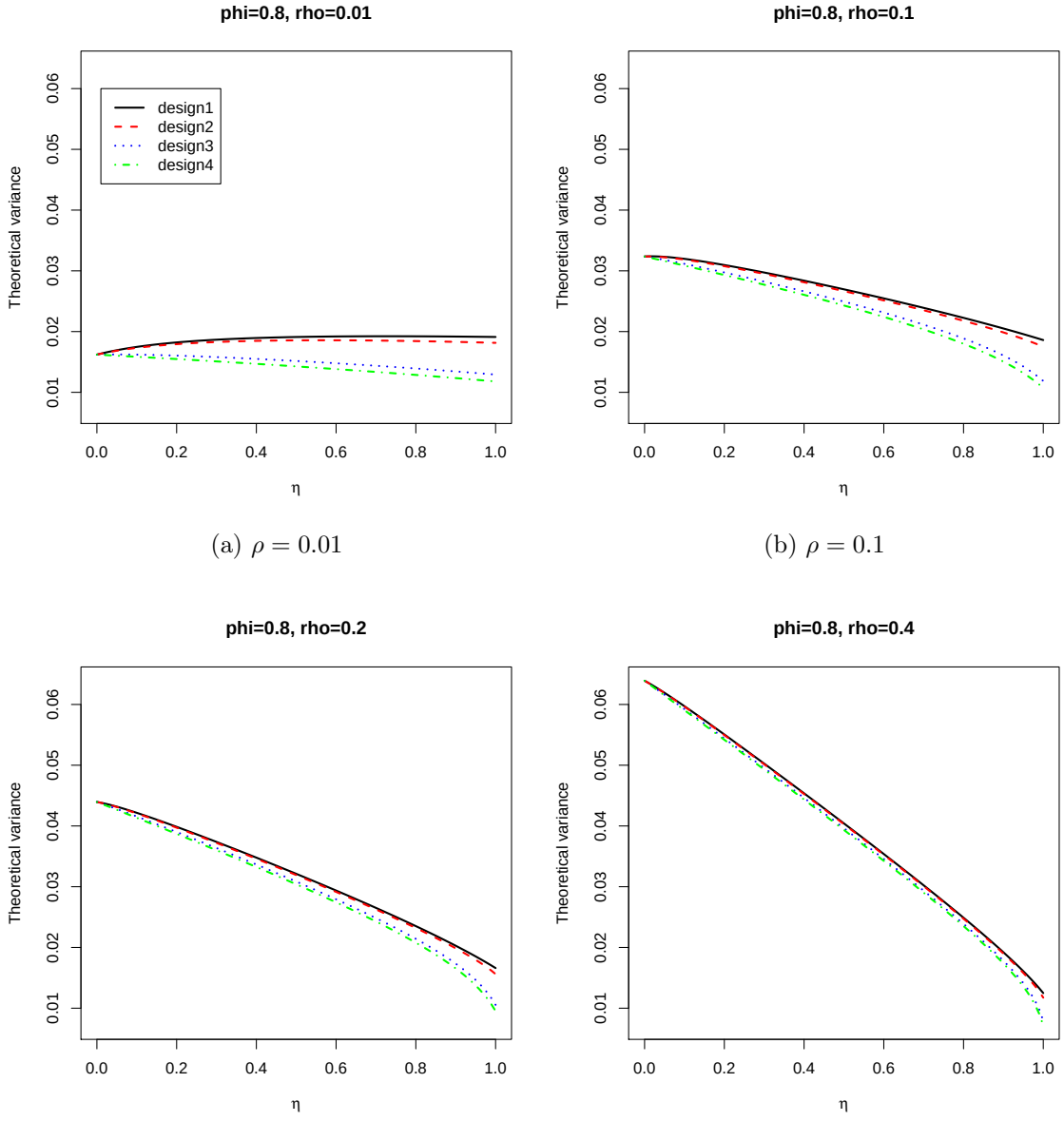
(a) $\rho = 0.01$

(b) $\rho = 0.1$

(c) $\rho = 0.2$

(d) $\rho = 0.4$

Figure B.5: $\text{Var}(\hat{\theta})$ vs. η ($\phi_b = 0.5$) when model (4.1) is correctly specified.



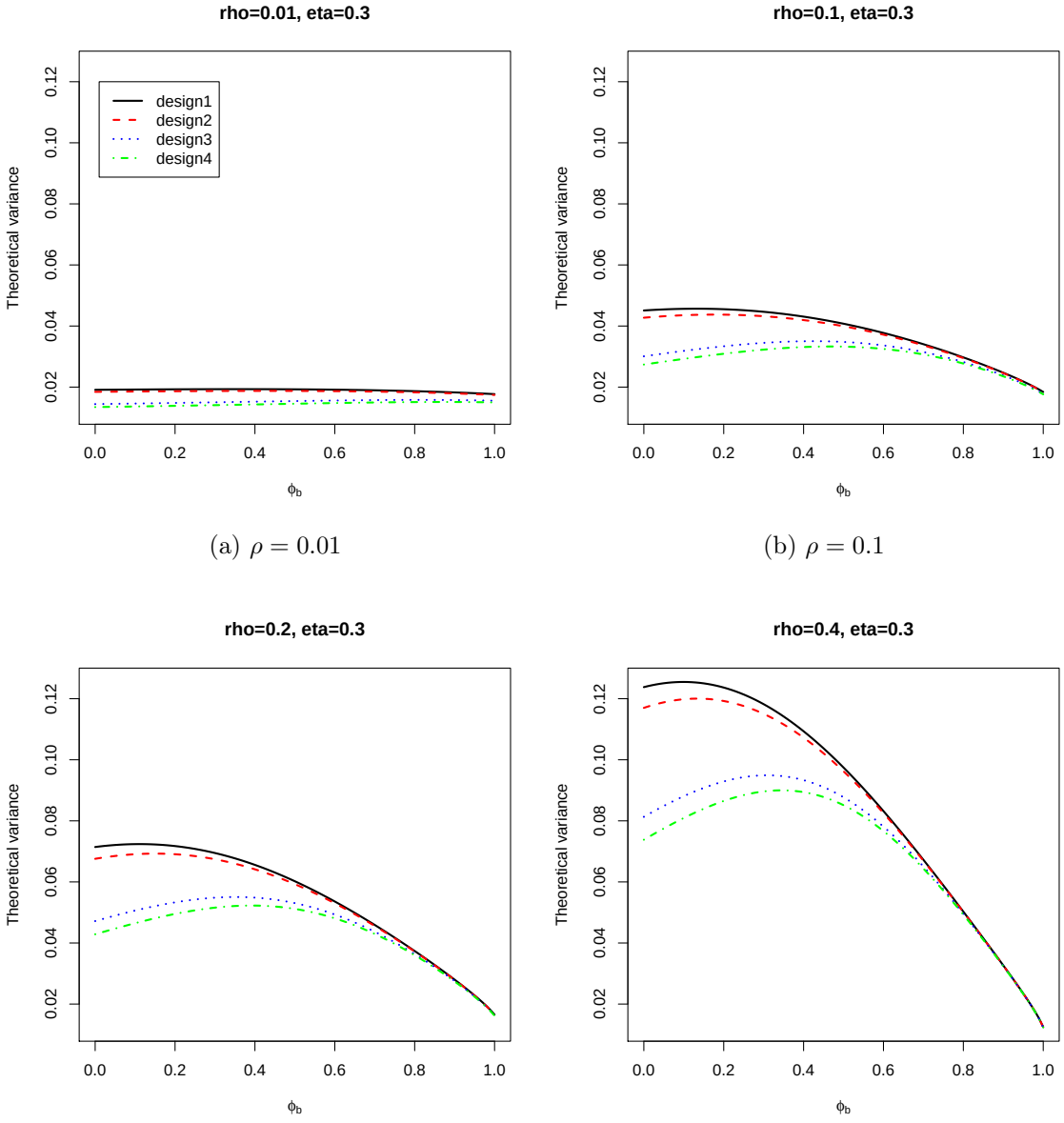
(a) $\rho = 0.01$

(b) $\rho = 0.1$

(c) $\rho = 0.2$

(d) $\rho = 0.4$

Figure B.6: $\mathrm{Var}(\hat{\theta})$ vs. η ($\phi_b = 0.8$) when model (4.1) is correctly specified.



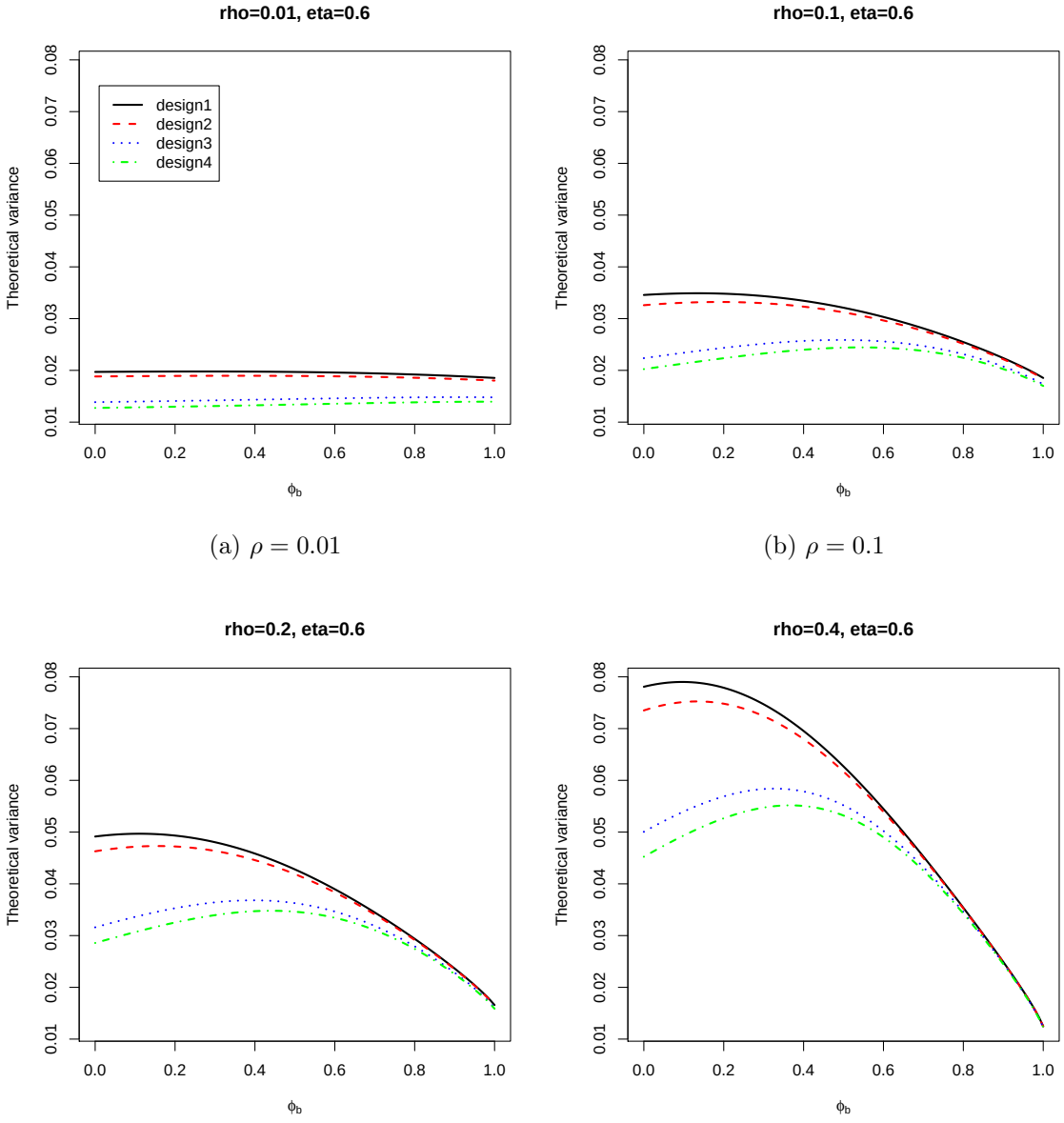
(a) $\rho = 0.01$

(b) $\rho = 0.1$

(c) $\rho = 0.2$

(d) $\rho = 0.4$

Figure B.7: $\text{Var}(\hat{\theta})$ vs. ϕ_b ($\eta = 0.3$) when model (4.1) is correctly specified.



(a) $\rho = 0.01$

(b) $\rho = 0.1$

(c) $\rho = 0.2$

(d) $\rho = 0.4$

Figure B.8: $\text{Var}(\hat{\theta})$ vs. ϕ_b ($\eta = 0.6$) when model (4.1) is correctly specified.

Bibliography

Agresti, A., Caffo, B., and Ohman-Strickland, P. Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis*, 47(3):639–653, 2004.

Amatya, A., Bhaumik, D., and Gibbon, R. D. Sample size determination for clustered count data. *Statistics in Medicine*, 32(24):4162–4179, 2013.

Baio, G., Copas, A., Ambler, G., Hargreaves, J., Beard, E., and Omar, R. Z. Sample size calculation for a stepped wedge trial. *BMC Trials*, 16(354), 2015.

Barker, D., McElduff, P., D'Este, C., and Campbell, M. J. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. *BMC Medical Research Methodology*, 16(69), 2016.

Bates, D., Mächler, M., Bolker, B., and Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

Bell, R. M., Ellickson, P. L., and Harrison, E. R. Do drug prevention effects persist into high school? how project alert did with ninth graders. *Prev Med*, 22(4): 463–483, 1993.

Bennett, P. N., Daly, R. M., Fraser, S. F., Haines, T., Barnard, R., Ockerby, C., and Kent, B. The impact of an exercise physiologist coordinated resistance exercise

program on the physical function of people receiving hemodialysis: a stepped wedge randomised control study. *BMC Nephrology*, 14(204), 2013.

Bland, J. M. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Medical Research Methodology*, 4(21), 2004.

Borm, G. F., Melis, R. J., Teerenstra, S., and Peer, P. G. Pseudo cluster randomization: a treatment allocation method to minimize contamination and selection bias. *Statistics in Medicine*, 24(23):3535–3547, 2005.

Brown, C., Hofer, T., Johal, A., Thomson, R., Nicholl, J., Franklin, B. D., and Lilford, R. J. An epistemology of patient safety research: a framework for study design and interpretation. Part 2. Study design. *Qualiy & Safety in Health Care*, 17(3):163–169, 2008.

Brown, C. A. and Lilford, R. J. The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology*, 6(54), 2006.

Bruce, M. L., Brown, E. L., Raue, P. J., Mlodzianowski, A. E., Meyers, B. S., Leon, A. C., et al. A randomized trial of depression assessment intervention in home health care. *Journal of the American Geriatrics Society*, 55(11):1793–1800, 2007.

Cambell, M. J., Donner, A., and Klar, N. Development in cluster randomized trials and statistics in medicine. *Statistics in Medicine*, 26(1):2–19, 2007.

Campbell, M. J. and Walters, S. J. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research.* Wiley, Chichester, 2014.

Casella, G. and Berger, R. L. *Statistical Inference*. Duxbury Press, Pacific Grove, CA, 2001.

Ciliberto, M. A., Sandige, H., Ndekha, M. J., Ashorn, P., Briend, A., Ciliberto, H. M., and Manary, M. J. Comparison of home-based therapy with ready-to-use therapeutic food with standard therapy in the treatment of malnourished malawian children: a controlled, clinical effectiveness trial. *The American Journal of Clinical Nutrition*, 81(4):864–870, 2005.

Cook, T. D. and Campbell, D. T. *Quasi-experimentation: Design and analysis issues for field settings*. Houghton Mifflin Company, MA, 1979.

Corey, D. M., Dunlap, W. P., and Burke, M. J. Averaging correlations: Expected values and bias in combined pearson rs and fisher's z transformations. *The Journal of General Psychology*, 125(3):245–261, 1998.

Diggle, P., Heagerty, P., and Y, L. K. *Analysis of longitudinal data*. Oxford University Press, NY, 2002.

Dimairo, M., Bradburn, M., and Walters, S. J. Sample size determination through power simulation; practical lessons from a stepped wedge cluster randomised trial (sw crt). *Trials*, 12(1):26, 2011.

Doig, G. S., Simpson, F., Finfer, S., Delaney, A., Davies, A. R., Mitchell, I., et al. Effect of evidence-based feeding guidelines on mortality of critically ill adults: a cluster randomized controlled trial. *Journal of the American Medical Association*, 300(23):2731–2741, 2008.

Donner, A. and Klar, N. Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. *American Journal of Epidemiology*, 140(3): 279–289, 1994.

Donner, A., Birkett, N., and Buck, C. Randomization by cluster: sample size requirements and analysis. *American Journal of Epidemiology*, 114(6):906–914, 1981.

Eldridge, S. and Kerry, S. *A Practical Guide to Cluster Randomised Trials in Health Services Research*. Wiley, United Kingdom, 2012.

Eldridge, S. M., Ashby, D., and Kerry, S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology*, 35(5):1292–1300, 2006.

Eldridgea, S. M., Ashbyb, D., Federa, G. S., Rudnickab, A. R., and Ukoumunne, O. C. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clinical Trials*, 1(1):80–90, 2004.

Feldman, H. A. and McKinlay, S. M. Cohort versus cross-sectional design in large filed trials: precision, sample size, and a unifying model. *Statistics in Medicine*, 13 (1):61–78, 1994.

Ferron, J., Dailey, R., and Yi, Q. Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research*, 37(3):379–403, 2002.

Fisher, R. A. *Statistical methods for research workers*. Oliver & Boyd, Edinburgh, 1958.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. *Applied Longitudinal Analysis*. Wiley, New Jersey, 2011.

Friedman, L. M., Furberg, C. D., and DeMets, D. *Fundamentals of Clinical Trials*. Springer, New York, 2010.

Gambia Hepatitis Study Group. The gambia hepatitis intervention study. *Cancer Research*, 47(21):5782–5787, 1987.

Gardiner, J. C., Luo, Z., and Roman, L. A. Fixed effects, random effects and gee: What are the differences? *Statistics in Medicine*, 28(2):221–239, 2009.

Giraudeau, B., Ravaud, P., and Donner, A. Sample size caculation for cluster randomized cross-over trials. *Statistics in Medicine*, 27(27):5578–5585, 2008.

Gruber, J. S., Reygadas, F., F, A. B., Ray, I., Nelson, K., and Colford, J. M. J. A stepped wedge, cluster-randomized trial of a household UV-Disinfection and safe storage drinking water intervention in rural Baja California Sur, Mexico. *The American Journal of Tropical Medicine and Hygiene*, 89(2):238–245, 2013.

Hardin, J., Garcia, S. R., and Golan, D. A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, 7(3):1733–1762, 2013.

Harville, D. A. *Matirx Algebra From A Statistician's Perspective*. Springer Science+Business Media, LLC, NY, 2012.

Hayes, R. and Bennet, S. Simple sample size calculation for cluster-randomization trials. *International Journal of Epidemiology*, 28(2):319–326, 1999.

Heagerty, P. J. and Kurland, B. F. Misspecified maximum likelihood estimates and generalized linear models. *Biometrika*, 88(4):973–985, 2001.

Hejblum, G., Chalumeau-Lemoine, L., Ioos, V., Boelle, P. Y., Salomon, L., Simon, T., Vibert, J.-F., and Guidet, B. Comparison of routine and on-demand prescription of chest radiographs in mechanically ventilated adults: a multicentre, cluster-randomised, two-period crossover study. *Lancet*, 374(9702):1687–1693, 2009.

Hemming, K., Girling, A. J., Stich, A. J., Marsh, J., and Lilford, R. J. Sample size calculations for cluster randomized controlled trials with a fixed number of clusters. *BMC Medical Research Methodology*, 11(102):102–112, 2011.

Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J., and Lilford, R. J. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *British Medical Journal*, 350, 2015a.

Hemming, K., Lilford, R., and Girling, A. J. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Statistics in Medicine*, 34(2):181–196, 2015b.

Heo, M. and Leon, A. C. Statistical power and sample size requirements for three level hierarchical cluster randomized trials. *Biometrics*, 64(4):1256–1262, 2008.

Heo, M. and Leon, A. C. Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials. *Statistics in Medicine*, 28(6):1017–1027, 2009.

Huang, X. Diagnosis of random-effect model misspecification in generalized linear mixed models for binary response. *Biometrics*, 65(2):361–368, 2009.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., Bruckner, T., and Satariano, W. A. To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, 21(4):467–474, 2010.

Hughes, J. P., Goldenberg, R. L., Wilfert, C. M., Valentine, M., Mwinga, K. G., Guay, L. A., Mmiro, F., and Stringer, J. S. A. Design of the hiv prevention trials network (HPTN) protocol 054: a cluster randomized crossover trial to evaluate combined access to nevirapine in developing countries. *Technical Report 195, University of Washington, Department of Biostatistics*, 2003.

Hussey, M. A. and Hughes, J. P. Design and analysis of stepped wedge cluster radomized trials. *Contemporary Clinical Trials*, 28(2):182–191, 2007.

Juul, L., Maindal, H. T., Zoffmann, V., Frydenberg, M., and Sandbaek, A. Effectiveness of a training course for general practice nurses in motivation support in type 2 diabetes care: a cluster-randomised trial. *PLoS One*, 9(5):e96683, 2014.

Kelly, J. A., St Lawrence, J. S., Diaz, Y. E., Stevenson, L. Y., Hauth, A. C., Brasfield, T. L., Kalichman, S. C., Smith, J. E., and Andrew, M. E. Hiv risk behavior reduction following intervention with key opinion leaders of population: an experimental analysis. *Am J Public Health*, 81(2):168–171, 1991.

Keogh-Brown, M. R., Bachmann, M. O., Shepstone, L., Hewitt, C., Howe, A., Ramsay, C. R., Song, F., Miles, J. N., Torgerson, D. J., Miles, S., Elbourne, D., Harvey, I., and Campbell, M. J. Contamination in trials of educational interventions. *Health Technology Assessment*, 11(43):iii, ix–107, 2007.

Killama, W. P., Tambatamba, B. C., Chintu, N., Rouse, D., Stringer, E., Bweupe, M., Yu, Y., and Stringer, J. S. Antiretroviral therapy in antenatal care to increase treatment initiation in HIV-infected pregnant women: a stepped-wedge evaluation. *AIDS*, 24(1):85–91, 2010.

Kim, H., Williamson, J. M., and Lyles, C. M. Sample-size calculations for studies with correlated ordinal outcomes. *Statistics in Medicine*, 24(19):2977–2987, 2005.

King, G., Gakidou, E., Ravishankar, N., Moore, R. T., Lakin, J., Vargas, M., Tellez-Rojo, M. M., Hernandez Avila, J. E., Hernandez Avila, M., and Hernandez Llamas, H. A "politically robust" experimental design for public policy evaluation, with application to the mexican universal health insurance program. *Journal of Policy Analysis and Management*, 26(3):479–506, 2007.

Kwok, O., West, S. G., and Green, S. B. The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A monte carlo study. *Multivariate Behavioral Research*, 42(3):557–592, 2007.

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

Liang, K. Y. and Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Litiere, S., Alonso, A., and Molenberghs, G. Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics*, 63(4):1038–1044, 2007.

Liu, S., Rovine, M. J., and Molenaar, P. C. M. Selecting a linear mixed model for longitudinal data: Repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychological Methods*, 17(1):15–30, 2012.

Lynn, P. *Methods of longitudinal surveys*. John Wiley & Sons Ltd, NJ, 2009.

Marsteller, J. A., Sexton, J. B., Hsu, Y. J., Hsiao, C. J., Holzmueller, C. G., Pronovost, P. J., et al. A multicenter, phased, cluster-randomized controlled trial to reduce central line-associated bloodstream infections in intensive care units. *Critical Care Medicine*, 40(11):2933–2939, 2012.

Matthews, J. and Altman, D. Analysis of serial measurements in medical research. *Britsh Medical Journal*, 300:230–235, 1990.

McCulloch, C. E. and Neuhaus, J. M. Misspecifying the shape of a random effects distribution: Why getting it wrong maynot matter. *Statistical SciencE*, 26(3): 388–402, 2011.

Mdege, N. D., Man, M.-S., Taylor, C. A., and Torgerson, D. J. Systematic review of stepped wedge cluster radomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology*, 64(9):936–948, 2011.

Murray, D. M., Perry, C. L., Griffin, G., Harty, K. C., Jacobs, D. R. J., Schmid, L., Daly, K., and Pallonen, U. Results from a statewide approach to adolescent tobacco use prevention. *Prev Med*, 21(4):449–472, 1992.

Pagoto, S. L., Schneider, K. L., Oleski, J., Bodenlos, J. S., Merriam, P., and Ma, Y. Design and methods for a cluster randomized trial of the sunless study: A skin

cancer prevention intervention promoting sunless tanning among beach visitors. *BMC Public Health*, 9(50), 2009.

Parienti, J. J. and Kuss, O. Cluster-crossover design: A method for limiting clusters level effect in community-intervention studies. *Contemporary Clinical Trials*, 28(3): 316–323, 2007.

Parienti, J.-J., du Cheyron, D., Ramakers, M., Malbruny, B., Leclercq, R., Le Coutour, X., and Charbonneau, P. Alcoholic povidone-iodine to prevent central venous catheter colonization: A randomized unit-crossover study. *Crit Care Med*, 32(3): 708–713, 2004.

Petersen, K. B. and Pedersen, M. S. *The Matrix Cookbook*. Technical University of Denmark, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

Reich, N. G. and Milstone, A. M. Improving efficiency in cluster-randomized study design and implementation: taking advantage of a crossover. *Open Access Journal of Clinical Trials*, 2014(6):11–15, 2014.

Reich, N. G., Myers, J. A., Obeng, D., Milstone, A. M., and Perl, T. M. Empirical power and sample size calculations for cluster-randomized and cluster-randomized crossover studies. *PLoS One*, 7(4):e35564, 2012.

Rhoads, C. H. The implications of "contamination" for experimental design in education. *Journal of Educational and Behavioral Statistics*, 36(1):76–104, 2011.

Rutterford, C., Copas, A., and Eldridge, S. Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, 44(3):1–17, 2015.

Scales, D. C., Golan, E., Pinto, R., Brooks, S. C., Chapman, M., Dale, C. M., et al. Improving appropriate neurologic prognostication after cardiac arrest. A stepped wedge cluster randomized controlled trial. *American Journal of Respiratory and Critical Care Medicine*, 194(9):1083–1091, 2016.

Schochet, P. Z. Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, 33(1):62–87, 2008.

Shih, W. Sample size and power calculations for periodotal and othe studies with clustered samples using the method of generalized estimating equations. *Biometrical Journal*, 39(8):899–908, 1997.

Simpson, J. M., Klar, N., and Donner, A. Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. *American Journal of Publiv Health*, 85(10):1378–1383, 1995.

Simunovic, M., Goldsmith, C., Thabane, L., McLeod, R., DeNardi, F., Whelan, T. J., and Levine, M. N. The quality initiative in rectal cancer (qirc) trial: study protocol of a cluster randomized controlled trial in surgery. *BMC Surgery*, 8(4), 2008.

Teenrenstra, S., Eldridge, S., Graff, M., de, H. E., and Borm, G. F. A simple sample size formula for analysis of covariates in cluster randomized trials. *Statististic in Medicine*, 31(20):2169–2178, 2012.

Teerenstra, S., Moerbeek, M., van Achterberg, T., Pelzer, B. J., and Borm, G. F. Sample size calculations for 3-level cluster randomized trials. *Clinical Trials*, 5(5): 486–495, 2008.

Tokola, K., Larocque, J., and Nevalainen, H. O. Power, sample size and sampling costs for clustered data. *Statistics and Probabiloty Letters*, 81(7):852–860, 2011.

Torgerson, D. J. Contamination in trials: is cluster randomisation the answer? *BMJ*, 322:355–357, 2001.

Turner, R. M., White, I. R., and Croudace, T. Analysis of cluster randomized cross-over trial data: A comparison of methods. *Staitistics in Medicine*, 26(2):274–289, 2007.

Verbeke, G. and Lesaffre, E. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91 (443):217–221, 1996.

Verbeke, G. and Lesaffre, E. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23(4):541–556, 1997.

White, H. Maximum likelihood estimation of misspecified models. *Econometrica*, 50 (1):1–25, 1982.

Whitehead, J. Sample size calculations for ordered categorical data. *Statistics in Medicine*, 12(24):2257–2271, 1993.

Woertman, W., de Hoop, E., Moerbeek, M., Zuidema, S. U., Gerritsen, D. L., and Teerenstra, S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology*, 66(7):752–758, 2013.

Wood, D. A., Kotseva, K., Connolly, S., Jennings, C., Mead, A., Jones, J., et al. Nurse-coordinated multidisciplinary, family-based cardiovascular disease prevention programme (EUROACTION) for patients with coronary heart disease and asymptomatic individuals at high risk of cardiovascular disease: a paired, cluster-randomised controlled trial. *Lancet*, 371(9629):1999–2012, 2008.