Cluster Randomized Trials with Treatment Noncompliance

Booil Jo*
Department of Psychiatry & Behavioral Sciences
Stanford University
Stanford, CA 94305-5795
booil@stanford.edu

Tihomir Asparouhov
Muthén & Muthén

Bengt O. Muthén
Graduate School of Education & Information Studies
University of California, Los Angeles

Nicholas S. Ialongo
Department of Mental Health
Johns Hopkins University

C. Hendricks Brown
Department of Epidemiology and Biostatistics
University of South Florida

Under Review

*The research of the first author was supported by NIMH and NIDA (MH066319, DA11796, MH066247, MH40859). The research of the third author was supported by NIAAA (AA00230). We thank participants of the Prevention Science Methodology Group for helpful feedback.

Cluster Randomized Trials with Treatment Noncompliance

Abstract

Cluster randomized trials (CRT) have been widely used in field experiments treating a cluster (or group) of individuals as the unit of randomization. This study focuses particularly on situations where CRT are accompanied by a common complication in field experiments, namely treatment noncompliance. In CRT, compliance behavior may be related not only to individual characteristics of study participants, but also to the environment of clusters individuals belong to. Therefore, analyses ignoring the connection between compliance and CRT may not provide valid results. Although randomized field experiments often suffer from both noncompliance and clustering of the data, these features have been studied as separate rather than concurrent problems. On the basis of Monte Carlo simulations, this study demonstrates how CRT and noncompliance may affect statistical inferences and how these two complications can be accounted for simultaneously. In particular, the effect of randomized intervention on individuals who abide by the intervention assignment (complier average causal effect: CACE) will be the focus of the study. For estimation of intervention effects considering both noncompliance and CRT, an ML-EM estimation method is employed.

Introduction

Individual-level randomization is not always possible in field experiments for practical or ethical reasons. Therefore, instead of standard randomized controlled trials, cluster randomized trials (CRT) have been widely used in practice, treating a cluster (or group) of individuals as the unit of randomization. For example, in primary care settings, the randomization unit is often a doctor or a clinic (e.g., Dexter et al., 1998), where a number of patients belong to each doctor. In school settings, the randomization unit is often a classroom, where a number of students belong to each teacher (e.g., Ialongo et al., 1999). In CRT settings, individuals belonging to the same cluster are likely to show resemblance due to various factors such as common environment (e.g., clinic, classroom) and common deliverer of intervention treatment (e.g., doctor, teacher). If resemblance in outcomes among individuals in each cluster is ignored in analyzing outcomes in CRT, standard errors are usually underestimated, which results in inflation of a type I error. That is, one may overestimate the significance of treatment effects, or falsely conclude that treatment effects are significant when they are not. Previous studies have shown how design and analysis strategies need to be adjusted for efficient and fair evaluation of treatment effects in CRTs (e.g., Donner & Klar, 1996; Murray, 1998; Raudenbush, 1997). Multilevel analysis techniques developed in various statistical frameworks (e.g., Aitkin & Longford, 1986; Goldstein, 1986; Longford, 1993; McCulloch, 1997; Muthén & Satorra, 1995; Raudenbush & Bryk, 2002; Raudenbush, Yang, & Yosef, 2000; Stiratelli, Laird, & Ware, 1984; Wong & Mason, 1985; Zeger & Karim, 1991) provide relevant tools for analyzing data accounting for nested data structures.

Another common complication in randomized studies is treatment noncompliance. Unlike in laboratory experiments or on-site randomized studies, compliance behavior is often hard to control in larger scale field experiments with intensive treatment regimes.

Study participants' compliance behavior can be associated with various factors such as background characteristics and motivation. For example, people who are highly motivated or have a special interest in the treatment will be more likely to comply with the treatment. In estimating intervention effects in this situation, intention to treat (ITT) analysis has been considered a gold standard, where randomized groups are compared regardless of compliance status. As secondary analyses to complement ITT analysis results, as-treated or per protocol analyses have been widely used, if the effect of treatments when actually received is of interest. However, these secondary analysis methods may yield seriously biased estimates of treatment effects. Given the questionable validity of these analyses, statistical methods such as CACE (complier average causal effect) estimation have been developed to better estimate treatment efficacy taking into account noncompliance (e.g., Angrist, Imbens & Rubin, 1996; Bloom, 1984; Frangakis & Rubin, 1999; Goetghebeur & Molenberghs, 1996; Hirano, Imbens, Rubin, & Zhou, 2000; Imbens & Rubin, 1997; Jo, 2002a; Little & Yau, 1998).

Further statistical challenges arise when CRTs are accompanied by noncompliance. What is interesting with compliance behavior in the CRT setting is that it may be influenced not only by individual characteristics, but also by characteristics of the cluster an individual belongs to. For example, in the Johns Hopkins University Preventive Intervention Research Center's (JHU PIRC) school intervention trial (Ialongo et al., 1999), which was used as a prototype for the Monte Carlo simulations reported in this study, the unit of randomization was a classroom. In particular, in the Family-School Partnership (FSP) intervention, poor compliance of parents was one of the major complications. Further, parents' compliance with the intervention activities substantially varied depending on the classroom their children belonged to. When compliance is defined as completion of at least two thirds of intervention activities, average compliance rate in each classroom ranged from 5% to 100%. One possible explanation for this phenomenon would be that some teachers communicated better with parents and/or were more eager to encourage parents. Another possibility is that children in some classrooms were more ready and willing to collaborate with their parents due to their teachers' characteristics and/or due to other classroom environment (e.g., proportion of aggressive children). Resemblance in compliance behavior in the same cluster unit is a unique, but common, problem that did not receive proper attention until Frangakis, Rubin, and Zhou (2002) demonstrated the possibility of estimating intervention effect accounting for both noncompliance and clustering within a Bayesian analysis framework. However, there is still little recognition among researchers that these two complications are often closely related, creating unique problems that cannot be handled without concurrent consideration of clustering and noncompliance.

This study intends to promote an understanding of CRT and noncompliance as simultaneous complications and to facilitate joint analyses that consider both problems. In particular, the effect of intervention treatments on individuals who would abide by the intervention assignment (CACE) will be the focus. The difference from the previous study (Frangakis et al., 2002) is that the current study puts its main emphasis on understanding the mechanisms in which simultaneous complications affect the quality of causal effect estimates and how serious the impact is, and in which conditions the impact is greater. Further, this study employs a maximum likelihood estimation method (built in Mplus version 4 and higher; Muthén & Muthén, 1998-2006), which is more accessible to many applied researchers. Not covered by this study is ITT analysis with these complications. A failure to consider both complications has different consequences in ITT analysis, which is dealt with in another manuscript in preparation (Jo, 2006). This paper is presented in the following order. First, the CACE estimation method is briefly reviewed. Second, analytical complications from the simultaneous presence of noncompliance and CRT are defined by extending the conventional notion of intraclass correlation. Third, various consequences of these complications are demonstrated using Monte Carlo simulation studies. Fourth, a joint modeling of CRT and compliance within a multilevel mixture analysis framework is presented. Finally, implications and limitations of the study are discussed.

Complier Average Causal Effect (CACE)

In randomized field experiments, it is not rare to have a situation where a substantial number of study participants fail to receive the assigned treatment. In this case, if the effect of treatment when actually received (i.e., efficacy) is of interest, both as-treated analysis, which considers treatment receipt but ignores treatment assignment status, and per-protocol analysis, which excludes individuals who failed to receive the treatment, have been commonly applied as ways of adjusting for noncompliance. However, the results of these analyses are not only hard to interpret as causal effects, but also subject to substantial bias (Robins & Greenland, 1994; Sheiner & Rubin, 1995). This study employs the CACE estimation method, which is considered a better alternative to as-treated or per-protocol analysis. The key advantage of the CACE estimation method is that causal effect is defined on the basis of individuals' potential outcomes under every treatment assignment status (Angrist et al., 1996; Frangakis & Rubin, 2002). Given that, causal interpretation is possible with clarified assumptions, and sensitivity analysis is feasible if these assumptions are likely to be violated.

Assuming binary treatment assignment (treatment or control) and binary treatment receipt (receive or not) status, Angrist et al. (1996) defined four potential compliance types. Compliers are individuals who receive treatment only if they are assigned to the treatment condition. Never-takers are individuals who do not receive the treatment even if they are assigned to the treatment condition. Defiers are individuals who do the opposite of what they are assigned to do. Always-takers are the individuals who always receive the treatment no matter which condition they are assigned to. Among these four potential types of individuals, emphasis is often given to compliers and the rest are considered noncompliers. The current study considers the two most common types, compliers and never-takers, which is the case when individuals assigned to the control condition do not have access to the treatment (e.g., JHU PIRC trial: Ialongo et al., 1999). Since there is only one type of noncomplier (i.e., never-takers), noncomplier will be used in this paper to refer to never-taker.

If compliance status is observed both in the treatment and control conditions, the causal effect of treatment can be estimated for any potential types of individuals defined by Angrist et al. (1996). However, in practice, compliance status is not observed completely (and is often unobserved among individuals assigned to the control condition), which complicates the estimation of causal effects given compliance types. In identifying causal effects of treatment assignment for compliers, having no defiers (the assumption of monotonicity: Imbens & Angrist, 1994) is crucial, whereas having noncompliers or always-takers is not. Along with the assumption of monotonicity, the assumption of the exclusion restriction (Angrist et al., 1996; Hirano et al., 2000; Imbens & Rubin, 1997; Jo, 2002a, b) plays a critical role in identifying CACE. Under this assumption, the effect of treatment assignment is allowed for compliers, but disallowed for always-takers and never-takers. For general discussion on underlying assumptions of CACE estimation, see, for example, Angrist et al. (1996), Jo (2002c), Little and Yau (1998), and West and Sagarin (2000).

Assuming only two compliance types (compliers and noncompliers) and the exclusion restriction, a continuous outcome Y for individual i ($i = 1, 2, 3,..., n$) can be expressed as

$$Y_i \; = \; \alpha_n \, n_i + \, \alpha_c \, c_i + \, \gamma_c \, c_i \, Z_i \; + \varepsilon_{ni} \, n_i + \varepsilon_{ci} \, c_i, \tag{1}$$

where $c_i = 0$ and $n_i = 1$ if individual i is a noncomplier, and $c_i = 1$ and $n_i = 0$ if individual i is a complier. The treatment assignment status $Z_i = 0$ if assigned to the control condition and $Z_i = 1$ if assigned to the treatment condition. The mean potential outcome when $Z = 0$ is α_n for noncompliers, and α_c for compliers. The average effect of treatment assignment for compliers is γ_c (CACE). A normally distributed residual for noncompliers is ε_{ni}, which has zero mean and the variance σ_n^2. A normally distributed residual for compliers is ε_{ci}, which has zero mean and the variance σ_c^2.

In the absence of covariates that predict compliance, the proportions of compliers and noncompliers can be expressed in the empty logistic regression as

$$\begin{aligned} P(c_i = 1) \; &= \; \pi_{ci}, \\ P(c_i = 0) \; &= \; 1 \; - \; \pi_{ci}, \\ logit(\pi_{ci}) \; &= \; \beta_0, \end{aligned} \tag{2}$$

where π_{ci} is the probability of being a complier for individual i, and β_0 is the logit intercept.

Based on Equations 1 and 2, three directly estimable population means can be expressed in terms of model parameters as

$$\mu_{1n} \;\; = \;\; \alpha_n, \tag{3}$$

$$\mu_{1c} \;\; = \;\; \alpha_c \, + \, \gamma_c, \tag{4}$$

$$\mu_0 \;\; = \;\; (1 - \pi_c) \, \alpha_n + \, \pi_c \, \alpha_c, \tag{5}$$

where π_c is the mean proportion of compliers in the population, which is directly estimable (i.e., there is a corresponding sample statistic) from the observed data assuming random assignment of treatments. The population mean potential outcome when $Z = 1$ is μ_{1n} for noncompliers and μ_{1c} for compliers. Both μ_{1n} and μ_{1c} are directly estimable from the observed data. The population mean potential outcome when $Z = 0$ is μ_{0n} (i.e., α_n) for noncompliers and μ_{0c} (i.e., α_c) for compliers. Under the assumption of the exclusion restriction, the effect of treatment assignment is disallowed for never-takers. Therefore, α_n is directly identified as μ_{1n} as shown in Equation 3.

Then, from Equations 3 and 5, α_c can be identified as

$$\alpha_c = \frac{\mu_0 - \pi_n \, \mu_{1n}}{\pi_c}. \tag{6}$$

From Equations 4 and 6, γ_c (CACE) can be identified as

$$\gamma_c = \mu_{1c} - \frac{\mu_0 - \pi_n \, \mu_{1n}}{\pi_c} = \frac{\mu_1 - \mu_0}{\pi_c}. \tag{7}$$

Under the condition that treatment assignment is random and that potential outcomes for each person are unrelated to the treatment status of other individuals (Stable Unit Treatment Value: Rubin, 1978, 1980, 1990), a large-sample based estimator of CACE is then formulated as

$$\gamma_c = \bar{y}_{1c} - \frac{\bar{y}_0 - p_n \, \bar{y}_{1n}}{p_c} = \frac{\bar{y}_1 - \bar{y}_0}{p_c}, \tag{8}$$

where $\bar{y}_{1c}$ is the sample mean outcome of the treatment group compliers, $\bar{y}_{1n}$ is the sample mean outcome of the treatment group never-takers, $\bar{y}_1$ is the sample mean outcome of the treatment group, $\bar{y}_0$ is the sample mean outcome of the control group, and p_c is the proportion of compliers in the treatment condition.

CACE in the CRT Setting

Let us assume a CRT where some study participants do not comply with the given treatment. Individual i ($i = 1, 2, 3,..., m_j$) now belongs to cluster j ($j = 1, 2, 3,..., G$). The expression in Equation 1 is modified as

$$Y_{ij} \;=\; \alpha_n \, n_{ij} + \alpha_c \, c_{ij} + \gamma_c \, c_{ij} \, Z_j \; + \varepsilon_{nbj} \, n_{ij} + \varepsilon_{nwij} \, n_{ij} \; + \varepsilon_{cbj} \, c_{ij} + \varepsilon_{cwij} \, c_{ij}, \qquad (9)$$

where Z_j denotes the cluster-level randomization status. The macro-unit residuals ε_{nbj} (noncompliers) and ε_{cbj} (compliers) represent cluster-specific effects given Z, and are assumed to be normally distributed with zero mean and the between-cluster variances σ^2_{nb} (noncompliers) and σ^2_{cb} (compliers). The micro-unit residuals ε_{nwij} (noncompliers) and ε_{cwij} (compliers) are assumed to be normally distributed with zero mean and the within-cluster variance σ^2_{nw} (noncompliers) and σ^2_{cw} (compliers), which are equal across clusters. The total residual variance is the sum of the between- and within-cluster variances.

The logistic regression in Equation 2 is modified as

$$\begin{aligned} P(c_{ij} = 1) \;&=\; \pi_{cij}, \\ P(c_{ij} = 0) \;&=\; 1 \;-\; \pi_{cij}, \\ logit(\pi_{cij}) \;&=\; \beta_0 + \varepsilon_{cj}. \end{aligned} \qquad (10)$$

where the between-cluster residual ε_{cj} has zero mean and a variance of ζ^2_b. The logit value varies across clusters ($\beta_0 + \varepsilon_{cj}$), meaning that the proportion of compliers differs across clusters.

Defining Complications in CRT with Noncompliance

In cluster randomized trials, inflation of variance is usually expected due to similarity among individuals in the same cluster. If data are appropriately analyzed considering inflation of variance, the resulting statistical power is usually lower than that in trials with individual-level randomization. If data are analyzed ignoring inflation of variance, the resulting type I error rate will be incorrectly inflated. Intraclass correlation (ICC) has been commonly used to gauge possible inflation of variance in CRTs.

Conventionally, ICC represents the level of resemblance among individuals belonging to the same cluster in terms of outcomes. However, ICC defined in this way may not well represent resemblance among individuals in CRTs accompanied by noncompliance. That is, individuals in the same cluster are likely to show resemblance not only in terms of outcomes, but also in terms of compliance. Further, the level of resemblance in outcomes may vary across different compliance types. In this case, the conventional ICC may not be informative in evaluating the impact of clustering on type I error rates. Given that, the definition of ICC is extended in this study to properly reflect situations where both clustering and noncompliance are present.

Outcome Intraclass Correlation

In the presence of noncompliance, outcome ICC can differ across compliance types. For example, in the FSP intervention condition in the JHU PIRC trial, when compliance is defined as completion of at least two thirds of intervention activities, the ICC estimate among noncompliers was much higher (0.25) than that of compliers (0.05) in terms of the shy behavior outcome at Grade 2. In other words, children's shy behavior outcome was more sensitive to teacher or classroom environment when parents complied poorly with the intervention activities, which may be interpreted as an indicator of low level parental involvement in general. The overall ICC (0.13) does not reflect this heterogeneity in resemblance across compliers and noncompliers.

From Equation 9, the intraclass correlation coefficient in outcome Y for noncompliers given Z is

$$\text{ICC}_{Yn} = \frac{\sigma^2_{nb}}{\sigma^2_{nb} + \sigma^2_{nw}}, \tag{11}$$

where σ^2_{nb} denotes the between-cluster variance and σ^2_{nw} denotes the within-cluster variance for noncompliers given Z. The total variance is the sum of the between- and within-cluster variances ($\sigma^2_n = \sigma^2_{nb} + \sigma^2_{nw}$).

The intraclass correlation coefficient in outcome Y for compliers given Z is

$$\text{ICC}_{Yc} = \frac{\sigma^2_{cb}}{\sigma^2_{cb} + \sigma^2_{cw}}, \tag{12}$$

where σ^2_{cb} denotes the between-cluster variance and σ^2_{cw} denotes the within-cluster variance for compliers given Z. The total variance is the sum of the between- and within-cluster variances ($\sigma^2_c = \sigma^2_{cb} + \sigma^2_{cw}$).

Compliance Intraclass Correlation

In CRTs, not only outcomes, but also compliance can be similar among individuals in the same cluster. Consequently, the compliance rate may vary across different clusters. There are several ways to present heterogeneity across clusters in proportions (Agresti, 1990; Commenges & Jacqmin, 1994; Haldane, 1940; McCullagh & Nelder, 1989; Snijder & Bosker, 1999). In line with McKelvey and Zavoina (1975), the intraclass correlation coefficient in compliance can be defined from Equation 10 as

$$\text{ICC}_C = \frac{\zeta^2_b}{\zeta^2_b + \pi^2/3\,,} \tag{13}$$

where ζ_b^2 is the between-cluster variance and $\pi^2/3$ is the variance for the within-cluster residual in the logistic distribution. ICC$_C$ represents the degree of resemblance in compliance among individuals belonging to the same cluster. For example, in the FSP intervention condition in the JHU PIRC trial, the ICC$_C$ estimate is 0.37, which reflects a substantial variation in the average compliance rate across classrooms (average compliance rate ranged from 5% to 100%).

Consequences of Simultaneous Complications

Intraclass correlations shown in Equations 11, 12, and 13 raise some new questions that used to be irrelevant in the CRT setting until we started considering noncompliance, such as 1) whether ICC$_C$ alone has any impact on variance misestimation (inflation of the type I error rate), 2) whether different combinations of ICC$_{Yn}$ and ICC$_{Yc}$ differently influence variance misestimation, and 3) whether the impact of ICC$_{Yn}$ and ICC$_{Yc}$ varies depending on the level of ICC$_C$. These speculations, if they turn out to be true, will provide compelling reasons for the simultaneous consideration of CRT and noncompliance in the analysis. This section explores various settings to examine which of these interactions between noncompliance and CRT have actual impact on variance misestimation, and in what conditions the impacts are more substantial. Monte Carlo simulations are employed for this purpose, since it is not straightforward to analytically derive possible inflation of the type I error rate, given missing compliance information and mixture distributions of different compliance types.

Data Generation

The Monte Carlo simulation results presented in this study are based on 500 replications. The size of each cluster (m) is either 20 or 40, and the total number of clusters (G) is 100 (50 in the control and 50 in the treatment condition). Although simulation settings are mostly based on the JHU PIRC FSP school intervention trial, a larger number of clusters (100 in this study compared to 18 in the JHU Study) is employed to avoid another source of variance misestimation and to focus on variance misestimation only due to intraclass correlations. The true ratio of the treatment and control groups is 50%:50% and the true compliance rate is 50% in all simulation settings.

Outcome ICC values are decided on the basis of the JHU Study ($\mathrm{ICC}_{Yn} = 0.25$ and $\mathrm{ICC}_{Yc} = 0.05$ for the shy behavior outcome at Grade 2). To examine the impact of different outcome ICC compositions, four different combinations of outcome ICC values are considered. They are 1) when neither noncompliers nor compliers have any outcome ICC ($\mathrm{ICC}_{Yn} = 0.0$ and $\mathrm{ICC}_{Yc} = 0.0$), 2) when only noncompliers have a substantial outcome ICC ($\mathrm{ICC}_{Yn} = 0.2$ and $\mathrm{ICC}_{Yc} = 0.0$), 3) when only compliers have a substantial outcome ICC ($\mathrm{ICC}_{Yn} = 0.0$ and $\mathrm{ICC}_{Yc} = 0.2$), and 4) when noncompliers and compliers have the same moderate level of ICC ($\mathrm{ICC}_{Yn} = 0.1$ and $\mathrm{ICC}_{Yc} = 0.1$). The first setting ($\mathrm{ICC}_{Yn} = \mathrm{ICC}_{Yc} = 0.0$) is considered to examine whether there is a pure impact of ICC_C in the absence of outcome ICC.

The true compliance ICC value ranges from 0.0 to 1.0 (In the JHU Study, ICC_C was about 0.37 in the intervention condition). The zero ICC_C indicates that compliance behavior is independent of the clusters individuals belong to. This setting is considered to examine whether different compositions of ICC_{Yn} and ICC_{Yc} have different impact on variance inflation and the subsequent variance misestimation in the absence of ICC_C. The perfect ICC_C (i.e., 1.0) is the other extreme situation, where every individual in the same cluster shows the same compliance behavior. This could be a possible scenario depending on how compliance is decided and how intervention treatments are delivered. For example, if a teacher or a doctor, who represents the unit of randomization, delivers the intervention and if study participants do not have much room for independent decision on compliance, it is likely that compliance is decided at the cluster level.

Another key component of the simulation settings is the distributional distance between compliers and noncompliers. Given missing compliance information, precision of the CACE estimate depends on how well the mixtures of distributions are separated. Therefore, the distance between the two groups normally improves the estimation quality (i.e., the farther apart the distributions, the better the precision; Jo, 2002c). In the CRT setting, however, having a farther distance does not necessarily have a positive impact on variance estimation. To represent the distance between the two distributions, three conditions are considered. Given Z, noncompliers and compliers are 1) 0.0 SD (standard deviation) apart, 2) 0.5 SD apart, or 3) 1.0 SD apart. In the JHU Study, noncompliers and compliers were approximately 1.0 SD apart in the FSP intervention condition.

Data were generated according to Equations 9 to 10. Specifically, the true within-cluster variances σ^2_{nw} and σ^2_{cw} take values of 1.0, 0.9, and 0.8. The true between-cluster variances σ^2_{nb} and σ^2_{cb} take values of 0.0, 0.1, and 0.2 to reflect ICC_{Yn} and ICC_{Yc} of 0.0, 0.1 and 0.2 given the total variance of 1.0. The true control condition noncomplier mean α_n is 1.0, and the true control condition complier mean α_c takes values of 1.0, 1.5, and 2.0 to reflect the distance between noncompliers and compliers (0.0, 0.5, and 1.0 SD apart). The true treatment assignment effect for compliers γ_c (i.e., CACE) is 0.6 (effect size of 0.6 on the basis of the total variance), and the true treatment assignment effect for noncompliers is zero (i.e., exclusion restriction holds). The true logit intercept

β_0 is zero (i.e., 50% compliance) and the true between-cluster compliance variance ζ_b^2 takes values of 0.00, 0.82, 2.19, 13.15, and 10000 on the logit scale to reflect ICC$_C$ of 0.0, 0.2, 0.4, 0.8 and 1.0 according to Equation 13.

CACE Estimation without Considering CRT

Data were analyzed on the basis of Equations 1 and 2, which represent the standard CACE model without considering the fact that randomization was done at the cluster level. As in the data generation step, the exclusion restriction is assumed.

The current study employs a maximum likelihood estimation approach. Given that compliance type cannot be observed in the control condition, the observed-data likelihood function based on treatment assignment ($Z = 1$: treatment condition, $Z = 0$: control condition) and observed treatment receipt status ($D = 1$: received, $D = 0$: not received) is

$$\begin{aligned} L(\theta \mid data) \quad \propto \quad & \prod_{i \subset \{Z_i=1, D_i=0\}} (1-\pi_c)\, f(y_i \mid \mu_{1n}, \sigma_n^2) \quad \times \prod_{i \subset \{Z_i=1, D_i=1\}} \pi_c\, f(y_i \mid \mu_{1c}, \sigma_c^2) \\ \times & \prod_{i \subset \{Z_i=0, D_i=0\}} [(1-\pi_c)\, f(y_i \mid \mu_{0n}, \sigma_n^2) + \pi_c\, f(y_i \mid \mu_{0c}, \sigma_c^2)], \end{aligned} \tag{14}$$

where $\theta = (\pi_n, \pi_c, \mu_{1n}, \mu_{1c}, \mu_{0n}, \mu_{0c}, \sigma_n^2, \sigma_c^2)$ is the set of parameters in the model, and $f(y_i \mid \mu, \sigma^2)$ denotes the probability density of a normal distribution with mean μ and variance σ^2, and π_c denotes the proportion of compliers in the population.

By maximizing the likelihood in Equation 14 with respect to the parameters of interest θ, ML estimates are obtained. The unknown compliance status in the control condition is handled as missing data via the EM algorithm (Dempster, Laird, & Rubin, 1977; Little & Rubin, 1987; McLachlan & Krishnan, 1997; Tanner, 1996). The E step computes the expected values of the complete-data sufficient statistics given data y and current parameter estimates θ. The M step computes the complete-data ML estimates with complete-data sufficient statistics replaced by their estimates from the E step. This procedure continues until it reaches optimal status. Parametric standard errors are computed from the information matrix of the ML estimator using both the first- and the second-order derivatives under the assumption of normally distributed outcomes. In the current study, ML-EM estimation of CACE was carried out by the M*plus* program version 4.1 (Muthén & Muthén, 1998-2006).

Impact of Compliance Intraclass Correlations

In CRTs, individuals in the same cluster may resemble each other not only in terms of outcomes, but also in terms of compliance behavior. In particular, the Monte Carlo simulation results presented in this section is based on a hypothetical setting, where individuals in the same cluster are similar in terms of compliance behavior, but not in terms of outcomes. Though quite unrealistic, this setting is important to consider to examine whether ICC_C alone has any impact on variance misestimation in the absence of outcome intraclass correlations, which is an intriguing question that has not been explicitly considered in CRT data analysis practice.

[Figure 1]

Panel (a) in Figure 1 shows how the coverage of the CACE estimate decreases as ICC_C increases. In the simulations for (a), the cluster size is 20. The nominal 95% confidence interval coverage rate is 0.95 (or nominal type I error rate is 0.05). A coverage rate below 0.95 indicates that standard error estimates do not appropriately reflect variance inflation due to clustering of compliance behavior. In other words, the level of significance of the treatment effect is overstated. When compliers and noncompliers have homogeneous distributions (0.0 SD apart), the coverage rate stays close to the nominal level irrespective of the change in ICC$_C$. When compliers and noncompliers are 0.5 SD apart, the coverage rate begins to be affected especially by high ICC$_C$ such as 0.8 and 1.0. When compliers and noncompliers are 1.0 SD apart, the coverage rate decreases substantially with moderate to high ICC$_C$.

The simulation results shown in Panel (b) in Figure 1 are based on the same settings as those of (a), except that the cluster size is 40 instead of 20. It is well known that the cluster size affects the magnitude of variance inflation due to outcome intraclass correlation (Donner, Birkett, & Buck, 1981; Kish, 1965). Panel (b) shows that the same rule applies to compliance intraclass correlation. As the cluster size increases from 20 to 40, the impact of ICC$_C$ becomes more prominent, resulting in substantial deterioration of the coverage rate even with low levels of ICC$_C$. As in (a), the coverage rate stays close to the nominal level irrespective of the change in ICC$_C$ when compliers and noncompliers have homegenous distributions.

The interesting phenomenon shown in Figure 1 can be explained by variance inflation in compliance, which consequently leads to variance inflation in parameters that involve outcome and compliance. However, as compliers and noncompliers have more homogeneous outcome distributions, variance inflation in compliance has less impact on variance inflation in those parameters. For example, in Equation 9, compliance status can be thought of as a between-cluster covariate (not fully observed) if ICC$_C = 1.0$. In this case, if compliers and noncompliers are homogeneous given Z (i.e., $\alpha_n - \alpha_c = 0$), variance inflation in compliance does not influence the between-cluster variance of the outcome. However, as α_n and α_c have some distance between them, the between-cluster compliance status actually produces between-cluster variance in the outcome. In analyses ignoring CRT, this variance inflation cannot be taken into account and consequently causes variance misestimation of key parameters such as CACE.

Impact of Outcome Intraclass Correlations

The Monte Carlo simulation results in this section are presented to show whether different combinations of ICC_{Yn} and ICC_{Yc} differently influence variance misestimation, and whether the impact of ICC_{Yn} and ICC_{Yc} varies depending on the level of ICC_C. As in Figure 1, the distance between compliers and noncompliers is included as one of the key factors that influence variance inflation. In Figures 2 and 3, the first outcome ICC pattern (i.e., $\mathrm{ICC}_{Yn} = \mathrm{ICC}_{Yc} = 0$) is included as a reference setting, where only the impact of ICC_C can be observed as in Figure 1. The other three patterns represent the overall outcome ICC of 0.1, and therefore should have the same influence on variance misestimation unless heterogeneity in the outcome ICC across compliers and noncompliers plays a role. The evidence of interaction between ICC_C and ICC_Y can be found when the impact of ICC_Y changes as ICC_C changes.

Panel (a) in Figure 2 shows how the coverage of the CACE estimate changes depending on ICC_{Yn}, ICC_{Yc}, and ICC_C when compliers and noncompliers have homogeneous distributions. The pure impact of the outcome ICC can be observed when $\mathrm{ICC}_C = 0$. It is shown that different patterns of ICC_Y have a similar impact on variance inflation, and the impact of ICC_Y on variance misestimation increases as ICC_C increases, although the change is not dramatic. As compliers and noncompliers are farther apart from each other as shown in (b) and (c), the coverage rate, in general, decreases faster as ICC_C increases. It is also shown that different patterns of ICC_Y show different coverage rates (i.e., have different impact on variance misestimation). It becomes clear in (c) that ICC_Y has the greatest impact on variance inflation when ICC_Y is concentrated among compliers (i.e., $\mathrm{ICC}_{Yn} = 0.0$, $\mathrm{ICC}_{Yc} = 0.2$) and the smallest impact when it is concentrated among noncompliers (i.e., $\mathrm{ICC}_{Yn} = 0.2$, $\mathrm{ICC}_{Yc} = 0.0$), though the overall ICC is the same.

[Figure 2]

The simulation results shown in Figure 3 are based on the same settings as those of Figure 2, except that the cluster size is 40 instead of 20. Figures 2 and 3 show similar trends, though the coverage rate change has, in general, steeper slopes across varying levels of ICC$_C$ when the cluster size is larger. It is also shown that the pattern of ICC$_Y$ matters more as the cluster size increases. That is, as the cluster size increases and compliers and noncompliers are farther apart from each other, ICC$_Y$ has a much greater impact on variance inflation when ICC$_Y$ is concentrated among compliers (i.e., ICC$_{Yn}$ $= 0.0$, ICC$_{Yc}$ $= 0.2$) than when it is concentrated among noncompliers (i.e., ICC$_{Yn}$ $= 0.2$, ICC$_{Yc}$ $= 0.0$) or evenly distributed across noncompliers and compliers (i.e., ICC$_{Yn}$ $= 0.1$, ICC$_{Yc}$ $= 0.1$). The results imply that the coverage rate may detriorate at an alarming rate in CRTs that employ large cluster sizes (e.g., 100 per cluster) even when compliers and noncompliers have moderate distributional differences.

[Figure 3]

The results reported in Figures 1, 2, and 3 are subject to change if the exclusion restriction assumption does not hold. However, including violation of the exclusion restriction as an additional factor results in high dimensional interactions among compliance ICC, outcome ICC, and the exclusion restriction. In other words, it is unlikely that consistent conclusions can be reached about the simultaneous impact of these three factors. Besides, violation of the exclusion restriction influences not only the standard error, but also point estimates of the CACE, further complicating the evaluation of the varaince estimation quality. One accessible way of checking sensitivity to violation of the exclusion restriction assumption is to use pretreatment covariates that are predictors of compliance (Jo, 2002a, 2002b). However, this method is not recommended for analyses of CRT data ignoring clustering, because both within- and between-cluster covariates lose predicting power as ICC_C increases. For example, if $\text{ICC}_C = 1.0$, within-cluster covariates, which have only within-cluster variances, cannot predict compliance (i.e., logit coefficients are zero), which is a cluster-level variable. Between-cluster covariates cannot predict compliance either because any kind of between-cluster variance cannot be taken into account in the analysis ignoring clustering. However, in the analyses simultaneously considering CRT and noncompliance, both within- and between-cluster covariates can be properly handled as predictors of compliance, alleviating the potential impact of the exclusion restriction violation.

Simultaneous Consideration of Clustering and Noncompliance

Simulation studies shown in the previous section emphasized the unique portion of variance inflation when CRTs are accompanied by noncompliance. To appropriately reflect this inflation, CRT and noncompliance need to be considered simultaneously in the analysis. On the basis of Monte Carlo simulations, this section demonstrates the joint analysis that considers both complications. Pretreatment covariates are added to data generation and analysis models employed in the previous section. As clustering of data is considered, between- and within-cluster covariates can be properly handled in the joint analysis given compliance and outcome ICCs. For example, in the JHU PIRC Trial, teacher characteristics or classroom environment (e.g., average level of aggression) can be treated as between-cluster or contextual-level variables, whereas students' baseline behavioral measures can be treated as within-cluster or individual-level variables.

Data Generation

In the presence of covariates, the expression in Equation 3 is modified as

$$\begin{aligned} Y_{ij} \;=\;& \alpha_n\, n_{ij} + \alpha_c\, c_{ij} + \gamma_n\, n_{ij}\, Z_j \,+\, \gamma_c\, c_{ij}\, Z_j \,+ \\ & \boldsymbol{\lambda}'_{nb}\, n_{ij}\, \mathbf{x}_{bj} \,+ \boldsymbol{\lambda}'_{nw}\, n_{ij}\, \mathbf{x}_{wij} +\, \boldsymbol{\lambda}'_{cb}\, c_{ij}\, \mathbf{x}_{bj} +\, \boldsymbol{\lambda}'_{cw}\, c_{ij}\, \mathbf{x}_{wij} \,+ \\ & \varepsilon_{nbj}\, n_{ij} + \varepsilon_{nwij}\, n_{ij} \,+ \varepsilon_{cbj}\, c_{ij} + \varepsilon_{cwij}\, c_{ij}, \qquad (15) \end{aligned}$$

where $\mathbf{x}_{bj}$ is a vector of between-cluster covariates and $\mathbf{x}_{wij}$ is a vector of within-cluster covariates. The vectors of logit coefficients $\boldsymbol{\lambda}_{nb}$ and $\boldsymbol{\lambda}_{nw}$ represent between- and within-cluster covariate effects on Y for noncompliers. The vectors of logit coefficients $\boldsymbol{\lambda}_{cb}$ and $\boldsymbol{\lambda}_{cw}$ represent between- and within-cluster covariate effects on Y for compliers. It is assumed that the effect of treatment assignment does not vary across different values of covariates (additivity: Jo, 2002a), and the main effect of treatment assignment for noncompliers γ_n (NACE: noncomplier average causal effect) is allowed. In other words, the exclusion restriction is not imposed.

In the presence of pre-treatment covariates, the probability that individual i in cluster j will comply (π_{cij}) varies depending on the influence of covariates. The logistic regression in Equation 4 is modified as

$$\begin{aligned} P(c_{ij} = 1 \mid \mathbf{x}_{bj}, \mathbf{x}_{wij}) \;&=\; \pi_{cij}, \\ P(c_{ij} = 0 \mid \mathbf{x}_{bj}, \mathbf{x}_{wij}) \;&=\; 1 \,-\, \pi_{cij}, \\ logit(\pi_{cij}) \;&=\; \beta_0 + \boldsymbol{\beta}'_{1b}\, \mathbf{x}_{bj} + \boldsymbol{\beta}'_{1w}\, \mathbf{x}_{wij} + \varepsilon_{cj}, \qquad (16) \end{aligned}$$

where the vector of logit coefficients $\boldsymbol{\beta}_{1b}$ indicates the level of association between compliance and between-cluster covariates, and the vector of logit coefficients $\boldsymbol{\beta}_{1w}$ indicates the level of association between compliance and within-cluster covariates. Nonzero variance of ε_{cj} (i.e., ζ_b^2) means that the proportion of compliers differs across clusters conditional on these covariates.

The Monte Carlo simulation results presented in this section are based on 500 replications. The size of each cluster (m) is 40, and the total number of clusters (G) is 100. Data were generated on the basis of Equations 15 and 16. The true within-cluster residual variances σ^2_{nw} and σ^2_{cw} are 0.9 and 0.8 respectively. The between-cluster residual variances σ^2_{nb} and σ^2_{cb} are 0.1 and 0.2, reflecting ICC$_{Y\,n}$ of 0.1 and ICC$_{Y\,c}$ of 0.2 given the total residual variance of 1.0. The control condition noncomplier mean α_n is 1.0, and the control condition complier mean α_c is 2.0. The treatment assignment effect for noncompliers γ_n (i.e., NACE: noncomplier average causal effect) is -0.2, and the treatment assignment effect for compliers γ_c (i.e., CACE) is 0.6. One covariate (X_1) with zero mean and within-cluster variance of 1.0 was generated to represent within-cluster covariates. Another covariate (X_2) with zero mean and between-cluster variance of 1.0 was generated to represent between-cluster covariates. For these covariates, the within-cluster regression coefficients are -0.1 for noncompliers ($\boldsymbol{\lambda}_{nw\,X_1}$) and -0.2 for compliers ($\boldsymbol{\lambda}_{cw\,X_1}$). The between-cluster regression coefficients are 0.1 for noncompliers ($\boldsymbol{\lambda}_{nb\,X_2}$) and 0.2 for compliers ($\boldsymbol{\lambda}_{cb\,X_2}$). The logit intercept β_0 is zero (50% compliance) and the between-cluster compliance residual variance ζ^2_b is 2.191 on the logit scale to reflect ICC$_C$ of 0.4 conditioning on covariates. Both within- and between-cluster logit coefficients ($\boldsymbol{\beta}_{1w\,X_1}$ and $\boldsymbol{\beta}_{1b\,X_2}$) are 0.7 (odds ratio of approximately 2.0).

CACE Estimation without Considering CRT

First, data were analyzed ignoring CRT as in the previous section. That is, the model used for data analysis can be described as

$$Y_i = \alpha_n\, n_i + \alpha_c\, c_i + \gamma_n\, n_i\, Z_i + \gamma_c\, c_i\, Z_i + \boldsymbol{\lambda}'_n\, n_i\, \mathbf{x}_i + \boldsymbol{\lambda}'_c\, c_i\, \mathbf{x}_i + \varepsilon_{ni}\, n_i + \varepsilon_{ci}\, c_i, \qquad (17)$$

$$logit(\pi_{ci}) = \beta_0 + \boldsymbol{\beta}'_1\, \mathbf{x}_i, \qquad (18)$$

where cluster-level parameters and variances are completely removed from the data generation model described in Equations 15 and 16. The vector of covariates $\mathbf{x}_i$ consists of the within- (X_1) and the between-cluster (X_1) covariates and both of them are treated as individual-level variables in the analysis. As in the data generation model, the exclusion restriction is relaxed (i.e., γ_n is freely estimated), relying on the additivity assumption and covariates that are good predictors of compliance. True values for noncomplier and complier residual variances (σ_n^2 and σ_c^2) are set at 1.0 in Table 1 by combining between- and within-cluster residual variances. For other parameters, true values in the data generation model were used, though between- and within-cluster paremters are not distinguished in the analysis (Mplus input and output available at: URL to be provided).

Table 1 shows the results from the analysis using the CACE model ignoring CRT. As in the CACE estimation presented in the previous section, an ML-EM estimation method is used. It is demonstrated that the average standard error (SE) estimates do not capture variance inflation of many parameters. In particular, the average SE estimate of CACE is about half of the standard deviation of CACE estimates from 500 replications (i.e., empirical SD), indicating poor estimation of the variance of the CACE estimate. As a result, the 95% confidence interval coverage rate of CACE is only 0.662. In other words, the type I error rate is 0.338, which is more than six times the nominal rate. Low coverage rates are also evident among other cluster-level parameters. In the logistic regression of compliance on covariates, bias is observed not only in SE estimates, but also in point estimates. Logit coefficient estimates for both the within- and the between-level covariate ($\boldsymbol{\beta}_{1X_1}$ and $\boldsymbol{\beta}_{1X_2}$) are noticeably biased by ignoring compliance ICC$_C$. In this example, covariates still have reasonable predicting power to support the identification of γ_n (NACE). The results, however, imply that relying on covariate information to relax the exclusion restriction can be risky in CRTs with potentially high ICC$_C$.

[Table 1]

CACE Estimation Considering CRT

Simulation results presented in Table 2 are based on the analysis using the same model used for data generation, described in Equations 15 and 16. To estimate this model, we employed a formal multilevel mixture analysis (Asparouhov & Muthén, 2006; Muthén, 2004). For ML-EM estimation of CACE in this framework, the Mplus program (Muthén & Muthén, 1998-2006) version 4.1 was used (Mplus input and output available at: URL to be provided). Estimation details are given in the Appendix.

Table 2 shows the results from the CACE analysis considering both clustering and noncompliance. Within-cluster residual variances (σ^2_{nw}, σ^2_{cw}) are now separated from the between-cluster variances, allowing for a better comparison between compliance types. In the logistic regression of compliance, the within-cluster covariate logit coefficient ($\boldsymbol{\beta}_{1w\,X_1}$) estimate is close to the true value with a reasonable coverage rate. It is also shown from the logit intercept (β_0) estimation that compliance type is estimated properly taking into account ICC$_C$. Fixed effect between-cluster parameters show reasonable coverage rates. In particular, the coverage rate of the CACE is 0.952 (compared to 0.662 in the anlysis ignoring CRT). Noncomplier and complier intercepts (α_n, α_c) also show good point estimates and coverage rates. In the logistic regression of compliance, the average between-cluster covariate logit coefficient ($\boldsymbol{\beta}_{1b\,X_2}$) estimate is close to the true value with a reasonable coverage rate. Given correctly estimated covariate effects on compliance ($\boldsymbol{\beta}_{1w\,X_1}$, $\boldsymbol{\beta}_{1b\,X_2}$), average treatment assignment effects for compliers (CACE) and for noncompliers (NACE) are more likely to be successfully separated in the analysis considering CRT than in the analysis ignoring CRT. Between-cluster residual variances (σ^2_{nb}, σ^2_{cb}) are separated from the within-cluster residual variances, revealing the fact that between-cluster variance is more concentrated among compliers given covariates and treatment assignmnet. However, these random effect between-cluster parameters show somewhat low coverage rates. The results imply that, unless large numbers of clusters are available, the level of significance of between-cluster random effects should be interpreted with caution. We expect that the quality of standard error estimates will improve as the number of clusters increases (the number of clusters is currently 50 per compliance type). And finally, in the logistic regression of compliance, between-cluster variation of compliance is captured by the between-cluster residual variance (ζ^2_b), indicating a sizable ICC$_C$ given covariates.

[Table 2]

Conclusions

This study demonstrated the impact of intraclass correlations on variance inflation in the estimation of CACE in diverse CRT settings. The Monte Carlo simulation results showed various types of variance inflation that are unique to CRTs accompanied by treatment noncompliance. First, it was demonstrated that compliance intraclass correlation (ICC$_C$) itself can cause serious variance inflation in trials where cluster memebership is likely to influence individuals' compliance behavior. Second, given the same overall outcome intraclass correlation, the impact of clustering may differ substantially depending on how the intraclass correlation is spread across compliance types. Further, the impact of outcome intraclass correlations (ICC$_{Yn}$, ICC$_{Yc}$) may vary depending on the level of ICC$_C$. Whether compliers and noncompliers are similar or different in the control condition turned out to be an important factor that affects variance inflation. The higher the level of heterogeneity across complier and noncomplier distributions, the larger the impact of compliance and outcome intraclass correlations on variance inflation. The Monte Carlo simulation results indicate that the conventional intraclass correlation, which is limited to outcome with no distinction between compliers and noncompliers, may not be a sufficient indicator of how serious variance inflation is in CRTs with noncompliance.

A formal multilevel analysis combined with the mixture analysis was suggested as a way of dealing with both data clustering and noncompliance. On the basis of the formal multilevel analysis approach, between- and within-cluster-level parameters can be explicitly specified and estimated. On the basis of the the mixture model approach, compliance-class-specific parameters, such as CACE, can be estimated, considering mixture distributions of compliers and noncompliers. The interaction between data clustering and noncompliance can be explicitly modeled in the analysis that combines the two approaches. Another useful feature of the joint analysis is that it allows a flexible modeling of covariate effects. Considering the complex influence of covariates may help in better understanding intervention mechanisms. The ML-EM estimation of the multilevel mixture models has been implemented in the Mplus program (Muthén & Muthén, 1998-2006), providing an accessible tool for a wide range of researchers.

On the basis of incorrectly estimated standard errors, gauging statistical power is meaningless. Therefore, power was not explicitly discussed in the study, though power is of great concern in planning randomized trials. However, once variance inflation is properly taken into account, as demonstrated in the multilevel mixture analysis that considers both CRT and noncompliance, statistical power regains its validity. How statistical power varies depending on various settings of CRTs with noncompliance remains as a topic for future study. Another major limitation of the current study is that the proposed multilevel mixture analysis tends to provide biased standard error estimates when the number of clusters is small (e.g., 18 in the JHU PIRC Trial), which is a well-known problem in multilevel modeling. To avoid an additional source of variance misestimation and to focus on variance misestimation only due to intraclass correlations, a large number of clusters (i.e., 100 clusters) was considered in this study. However, small numbers of clusters are often employed in psycho-social intervention trials, and therefore the subsequent variance inflation is an important matter to be resolved. Further investigation is needed in this area to provide methods to improve the quality of standard error estimates given small numbers of clusters.

Appendix: ML-EM Estimation of CACE in the Multilevel Mixture Model Framework

The observed-data two-level likelihood function is based again on the treatment assignment Z and the observed treatment receipt status D. The likelihood for the $j-$th cluster is

$$\begin{aligned}L_j(\theta \mid data) \;\propto \int \Bigg(& \prod_{i \subset \{Z_{ij}=1, D_{ij}=0\}} (1-\pi_{cij})\, f_n(y_{ij} \mid data, \varepsilon_{nbj}) \quad \times \\ & \prod_{i \subset \{Z_i=1, D_i=1\}} \pi_{cij}\, f_c(y_{ij} \mid data, \varepsilon_{cbj}) \times \\ & \prod_{i \subset \{Z_i=0, D_i=0\}} [(1-\pi_{cij})\, f_n(y_{ij} \mid data, \varepsilon_{nbj}) + \pi_{cij}\, f_c(y_{ij} \mid data, \varepsilon_{cbj})]\Bigg) \phi(\varepsilon_{cbj}, \varepsilon_{nbj}, \varepsilon_{cj}) d\varepsilon_{cbj} d\varepsilon_{nbj} d\varepsilon_{cj},\end{aligned}$$

where f_n and f_c are the probability density of a normal distribution

$$f_n(y_{ij} \mid data, \varepsilon_{nbj}) = Exp\Bigg(- \frac{(y_{ij} - \alpha_n - \gamma_n Z_{ij} - \lambda'_{nb} x_{bj} - \lambda'_{nw} x_{wij} - \varepsilon_{nbj})^2}{2\sigma^2_{nw}} \Bigg) / (\sqrt{2\pi}\sigma_{nw})$$

$$f_c(y_{ij} \mid data, \varepsilon_{cbj}) = Exp\Bigg(- \frac{(y_{ij} - \alpha_c - \gamma_c Z_{ij} - \lambda'_{cb} x_{bj} - \lambda'_{cw} x_{wij} - \varepsilon_{cbj})^2}{2\sigma^2_{cw}} \Bigg) / (\sqrt{2\pi}\sigma_{cw})$$

π_{cij} is the probability of compliance

$$\pi_{cij} = \frac{Exp(\beta_0 + \beta'_{1b} x_{bj} + \beta'_{1w} x_{wij} + \varepsilon_{cj})}{1 + Exp(\beta_0 + \beta'_{1b} x_{bj} + \beta'_{1w} x_{wij} + \varepsilon_{cj})}.$$

and $\phi(\varepsilon_{cbj}, \varepsilon_{nbj}, \varepsilon_{cj})$ is the joint density function for the random effects

$$\phi(\varepsilon_{cbj}, \varepsilon_{nbj}, \varepsilon_{cj}) = Exp\Big(- \varepsilon^2_{cbj}/(2\sigma^2_{cb}) - \varepsilon^2_{nbj}/(2\sigma^2_{nb}) - \varepsilon^2_{nbj}/(2\zeta^2_b)\Big) / ((2\pi)^{1.5} \sigma_{cb} \sigma_{nb} \zeta_b).$$

The total likelihood function is

$$L(\theta \mid data) = \prod_j L_j(\theta \mid data).$$

The parameters in the model are

$$\theta = (\alpha_c, \alpha_n, \sigma_{nw}, \sigma_{cw}, \sigma_{nb}, \sigma_{cb}, \zeta_b, \lambda_{nw}, \lambda_{cw}, \lambda_{nb}, \lambda_{cb}, \beta_0, \beta_{1w}, \beta_{1b}).$$

By maximizing the total likelihood function with respect to the parameters of interest θ, ML estimates are obtained. Numerical integration is used to approximate the integration over the between level random effects. The unknown compliance status in the control condition and the between level random effects are handled as missing data via the EM algorithm. Parametric standard errors are computed from the information matrix of the ML estimator using both the first- and the second-order derivatives under the assumption of normally distributed outcomes. In the current study, ML-EM estimation of CACE was carried out by the M*plus* program version 4.1 (Muthén & Muthén, 1998-2006).

References

Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley.

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies (with discussion). *Journal of Royal Statistical Society, Ser. A, 149*, 1-43.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*, 444-455.

Asparouhov, T. & Muthén, B. O. (2006). Multilevel mixture models. Forthcoming in Hancock, G. R., & Samuelsen, K. M. (Eds.). (2007). *Advances in latent variable mixture models.* Greenwich, CT: Information Age Publishing, Inc.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review, 8*, 225-246.

Commenges, D., & Jacqmin, H. (1994). The intraclass correlation coefficient: distribution-free definition and test. *Biometrics, 50*, 517-526.

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

Dexter, P., Wolinsky, F., Gramelspacher, G., Zhou, X. H., Eckert, G., Waisburd, M., & Tierney, W. (1998). Effectiveness of computer-generated reminders for increasing discussions about Advance Directives and completion of Advance Directives. *Annals of Internal Medicine, 128*, 102-110.

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

Donner, A., & Klar, N. (1996). Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology, 49*, 435-439.

Frangakis, C. E., & Rubin, D. B. (1999). Addressing complications of intent-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika, 86*, 365-379.

Frangakis, C. E. & Rubin, D. B. (2002) Principal stratification in causal inference. *Biometrics, 58*, 21-29.

Frangakis, C. E., Rubin, D. B., & Zhou, X. H. (2002). Clustered encouragement design with individual noncompliance: Bayesian inference and application to advance directive forms. *Biostatistics, 3, 147-164.*

Goetghebeur, E., & Molenberghs, G. (1996). Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association, 91*, 928-934.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43-56.

Haldane, J. B. S. (1940). The mean and variance of χ_2, when used as a test of homogeneity, when expectations are small. *Biometrika, 31*, 346-355.

Hirano, K., Imbens, G. W., Rubin, D. B. & Zhou, X. H. (2000). Assessing the Effect of an Influenza Vaccine in an Encouragement Design. *Biostatistics, 1*, 69-88.

Ialongo, N. S., Werthamer, L., Kellam, S. G., Brown, C. H., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression and antisocial behavior. *American Journal of Community Psychology, 27*, 599-642.

Imbens, G. W., & Angrist. J. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62*, 467-476.

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with non-compliance. *The Annals of Statistics, 25*, 305-327.

Jo, B. (2002a). Estimating intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics, 27*, 385-420 (with discussion).

Jo, B. (2002b). Model misspecification sensitivity analysis in estimating causal effects of interventions with noncompliance. *Statistics in Medicine, 21*, 3161-3181.

Jo, B. (2002c). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods, 7*, 178-193.

Jo, B. (2006). Statistical power in intention-to-treat analysis in cluster randomized trials with noncompliance. Manuscript in preparation.

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

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Little, R. J. A., & Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3*, 147-159.

Longford, N. (1993). *Random Coefficient Models.* New York: Oxford University Press.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*, London: Chapman & Hall.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association, 92*, 162-170.

McKelvey, R. D., & Zavoina. W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4*, 103-120.

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions.* New York: Wiley.

Murray, D. M. (1998). *Design and Analysis of Group-Randomized Trials.* New York: Oxford University Press.

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345-368). Newbury Park, CA: Sage Publications.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological Methodology* (pp. 267-316). Cambridge, MA: Blackwell.

Muthén, L. K., & Muthén, B. O. (1998-2006). *Mplus user's guide.* Los Angeles: Muthén & Muthén.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173-185.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Thousand Oaks, CA: Sage.

Raudenbush, S. W., Yang, M. L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics, 9*, 141-157.

Robins, J. M., & Greenland, S. (1994). Adjusting for differential rates of PCP prophylaxis in high- versus low-dose AZT treatment arms in an AIDS randimized trial. *Journal of the American Statistical Association, 89*, 737-749.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6*, 34-58

Rubin, D. B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by D. Basu. *Journal of the American Statistical Association, 75*, 591-593.

Rubin, D. B. (1990). Comment on "Neyman (1923) and causal inference in experiments and observational studies." *Statistical Science, 5*, 472-480.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* Thousand Oaks, CA: Sage.

Sheiner, L. B., & Rubin, D. B. (1995). Intention to treat analysis and the goals of clinical trials. *Clinical Pharmacology and Therapeutics, 57*, 6-15.

Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics, 40*, 961-971.

Tanner, M. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributons and likelihood functions.* New York: Springer.

West, S. G., & Sagarin, B. J. (2000). Participant selection and loss in randomized experiments. In L. Bickman (Ed.), *Research Design* (pp. 117-154). Thousand Oaks, CA: Sage.

Wong, G. Y., & Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association, 80*, 513-524.

Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association, 86*, 79-86.

Table 1. Simulation: estimation of $CACE$ ignoring CRT (100 clusters, 40 per cluster)

Parameter	True Value	Average Estimate	Empirical SD	Average SE	95% CI Coverage
γ_n (NACE)	-0.200	-0.212	0.121	0.064	0.700
γ_c (CACE)	**0.600**	**0.608**	**0.132**	**0.064**	**0.662**
α_n	1.000	1.010	0.101	0.058	0.744
α_c	2.000	1.994	0.111	0.058	0.684
λ_{nX_1}	-0.100	-0.099	0.028	0.027	0.940
λ_{cX_1}	-0.200	-0.202	0.028	0.027	0.934
λ_{nX_2}	0.100	0.100	0.051	0.028	0.712
λ_{cX_2}	0.200	0.204	0.067	0.027	0.594
σ_n^2	1.000	0.996	0.047	0.040	0.918
σ_c^2	1.000	0.989	0.055	0.040	0.838
β_0	0.000	0.004	0.164	0.048	0.452
β_{1X_1}	0.700	0.515	0.055	0.048	0.066
β_{1X_2}	0.700	0.513	0.166	0.050	0.246

Table 2. Simulation: estimation of $CACE$ considering CRT (100 clusters, 40 per cluster)

Parameter	True Value	Average Estimate	Empirical SD	Average SE	95% CI Coverage
Within Level					
$\lambda_{nw\,X_1}$	-0.100	-0.099	0.026	0.026	0.952
$\lambda_{cw\,X_1}$	-0.200	-0.200	0.025	0.025	0.952
σ^2_{nw}	0.900	0.900	0.035	0.036	0.952
σ^2_{cw}	0.800	0.799	0.032	0.032	0.934
β_0	0.000	0.008	0.212	0.210	0.942
$\beta_{1w\,X_1}$	0.700	0.702	0.058	0.058	0.944
Between Level					
γ_n (NACE)	-0.200	-0.204	0.105	0.102	0.946
γ_c (CACE)	**0.600**	**0.603**	**0.119**	**0.123**	**0.952**
α_n	1.000	1.002	0.087	0.087	0.942
α_c	2.000	1.999	0.099	0.101	0.942
$\lambda_{nb\,X_2}$	0.100	0.099	0.048	0.047	0.936
$\lambda_{cb\,X_2}$	0.200	0.202	0.059	0.058	0.924
σ^2_{nb}	0.100	0.095	0.025	0.025	0.898
σ^2_{cb}	0.200	0.190	0.045	0.041	0.876
$\beta_{1b\,X_2}$	0.700	0.692	0.208	0.196	0.932
ζ^2_b	2.191	2.162	0.488	0.482	0.930

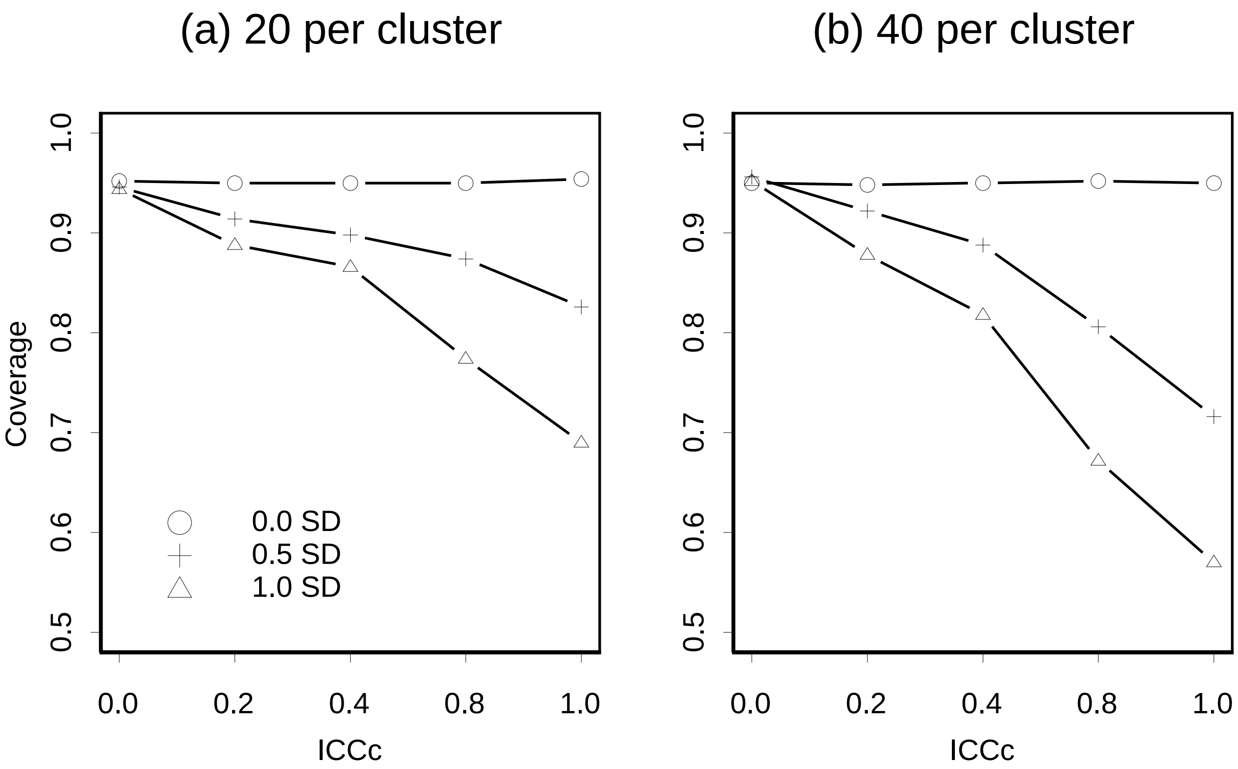

Figure 1: Impact of ICC_c on variance misestimation when $ICC_y = 0$ (i.e., $ICC_{yn} = ICC_{yc} = 0$) when complier and noncomplier means are 0.0, 0.5, and 1.0 standard deviation apart given treatment assignment.

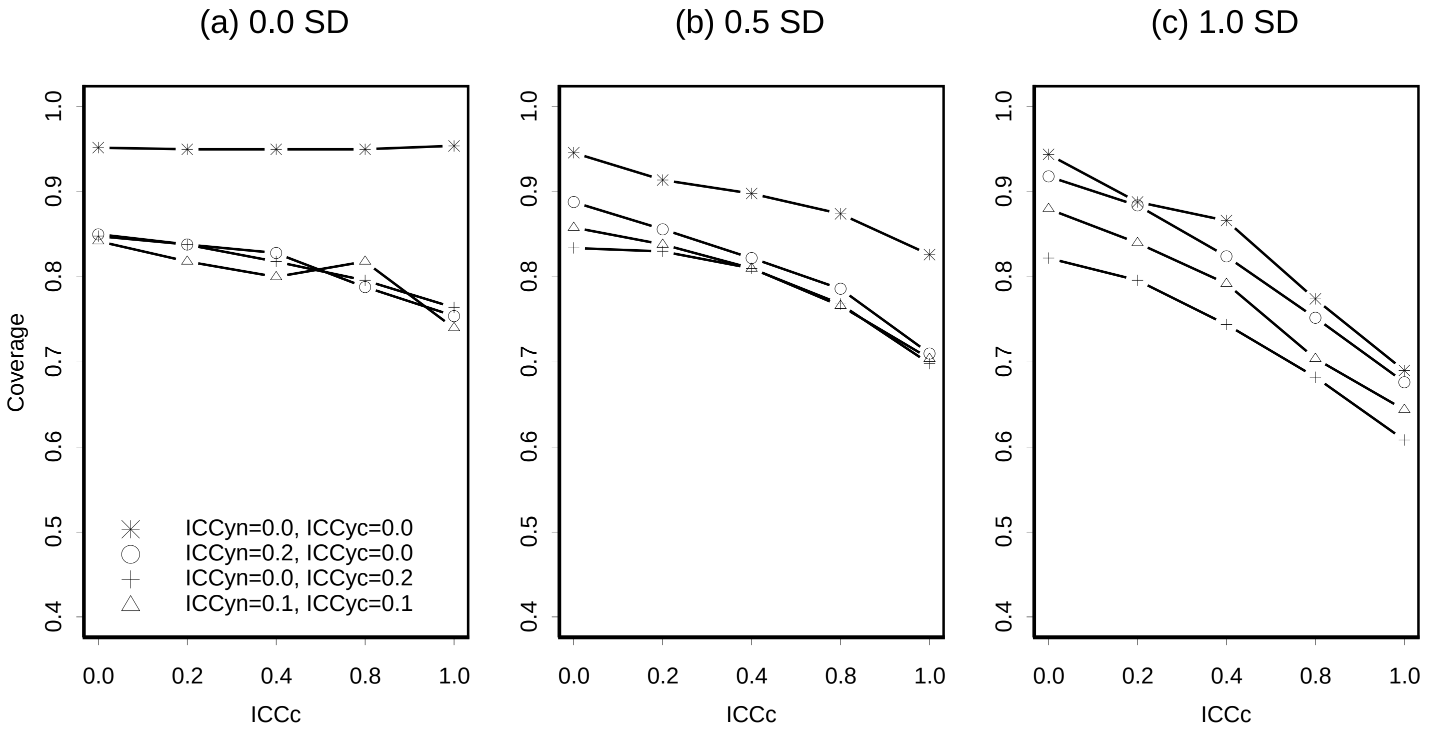

Figure 2: Impact of varying combinations of ICC_{yn}, ICC_{yc}, and ICC_c on variance misestimation when each cluster consists of 20 individuals. Complier and noncomplier means are (a) 0.0, (b) 0.5, and (c) 1.0 standard deviation apart given treatment assignment.

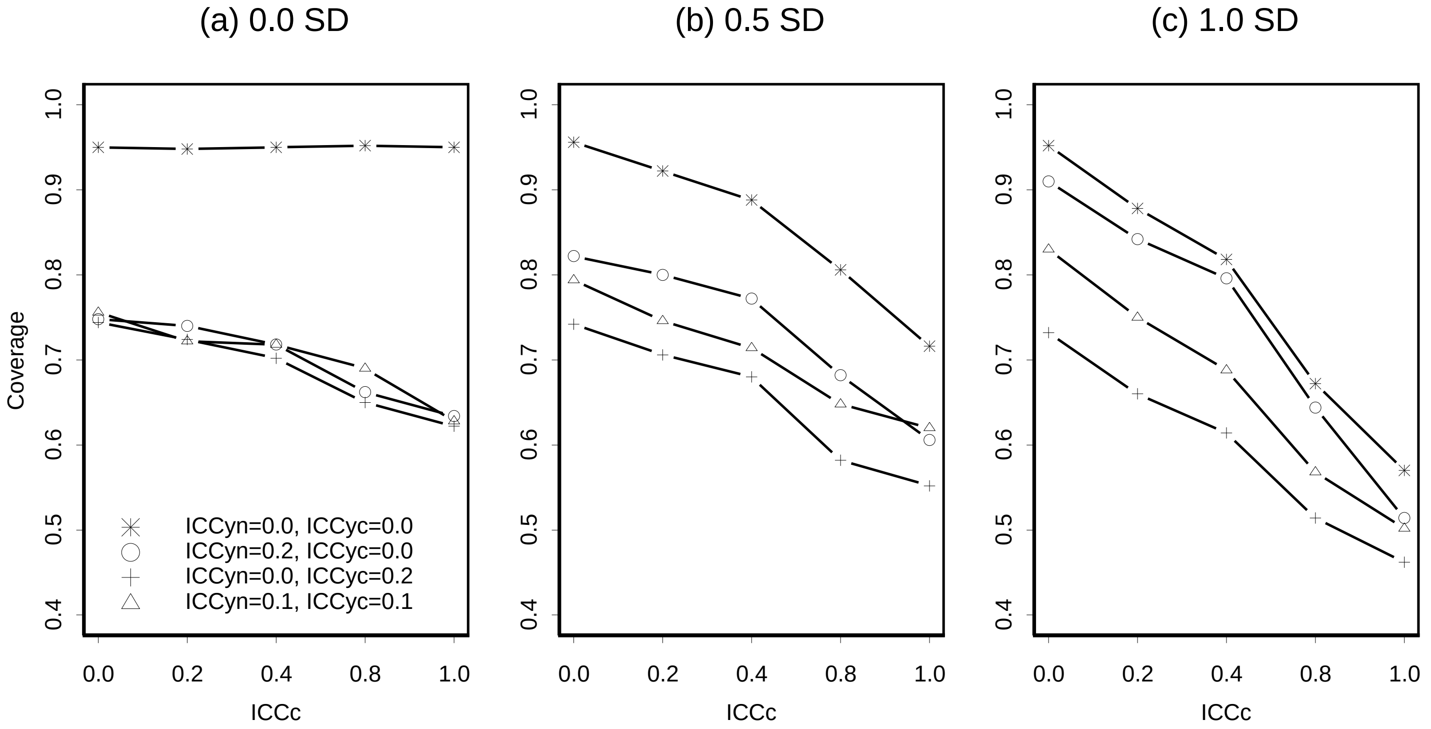

Figure 3: Impact of varying combinations of ICC_{yn}, ICC_{yc}, and ICC_c on variance misestimation when each cluster consists of 40 individuals. Complier and noncomplier means are (a) 0.0, (b) 0.5, and (c) 1.0 standard deviation apart given treatment assignment.