EVALUATION TECHNICAL ASSISTANCE UPDATE
for OAH & ACYF Teenage Pregnancy Prevention Grantees

December 2013 • Update 5

Frequently Asked Questions About The Implications of Clustering in Clustered Randomized Controlled Trials (RCTs)

As part of the technical assistance (TA) to TPP and PREIS grantees, the evaluation TA team will produce a series of updates that discuss topics relevant to the rigorous impact evaluations. Grantees' requests for TA and conversations with TA liaisons determine the topics and questions for these updates. This update answers frequently asked questions about the implications of clustering in clustered randomized controlled trials (RCTs).

What is a clustered RCT?

A *cluster* is a group of individuals or other clusters. For example, a classroom is a cluster of students, and a school is a cluster of classrooms. A *clustered RCT* is a randomized experiment in which clusters (as opposed to individuals) are randomly assigned to treatment and control groups.

What other words do people use to talk about clusters in data?

Researchers from different fields may use different terms to describe the same thing. For example, one researcher may describe data with clusters as *nested*; another may say it has a *hierarchical structure*; and another may call it *multilevel* data.

Why is it important to adjust for clustering in a clustered RCT?

When clusters are randomly assigned to treatment and control groups, the variance of the estimated impact is typically larger than when individuals are randomly assigned. If no adjustment is made for the effect of clustering, statistical significance of impact estimates will be overstated.

How does random assignment of clusters increase the variance of an impact estimate?

Randomly assigning clusters rather than individuals reduces the effective sample size of a study, leading to more variability in the impact estimate.

For example, consider a hypothetical experiment in which eight children from four families are randomly assigned to treatment and control groups, and the outcome of interest is the height of the children. Figure 1 shows the height (in inches) of each child, grouped by family.

If the eight *individual children* are randomly assigned to treatment and control groups, regardless of their families, then there are 70 ways to form a treatment and control group.[1] If there is no effect of being assigned to the treatment group, then the average impact (averaging across all 70 random assignments) is zero, and the variance of those impacts is 44 inches. Figure 2 shows a histogram representing the distribution of impacts across 70 random assignments.

If the four *families* (that is, clusters of children) are randomly assigned to treatment and control groups, then there are only six ways to form treatment and control groups. If there is no effect of being assigned to the treatment group, the average impact (averaging across all six random assignments) is zero, and the variance of those impacts is 100 inches. Figure 3 shows a histogram representing the distribution of impacts across six random assignments.

The difference in this example between randomly assigning individuals and randomly assigning clusters is stark. Although both approaches to random assignment yield the same expected (and unbiased) impact, the variance of the impact from these two approaches is very different, as is the shape of the impact distribution (it is much more "lumpy" when clusters are assigned).

Office of Adolescent Health

Figure 1. Height in Inches of Children in Four Families

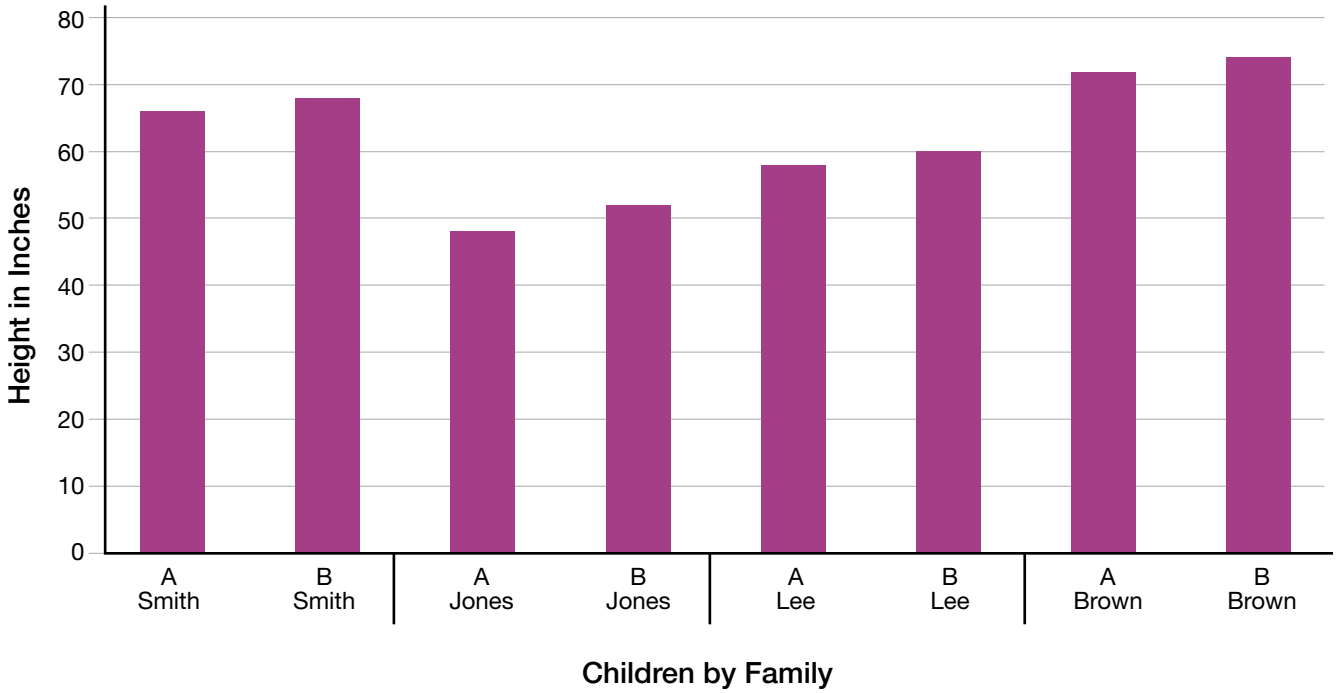


Figure 2. Histogram of Impacts When Children Are Randomly Assigned

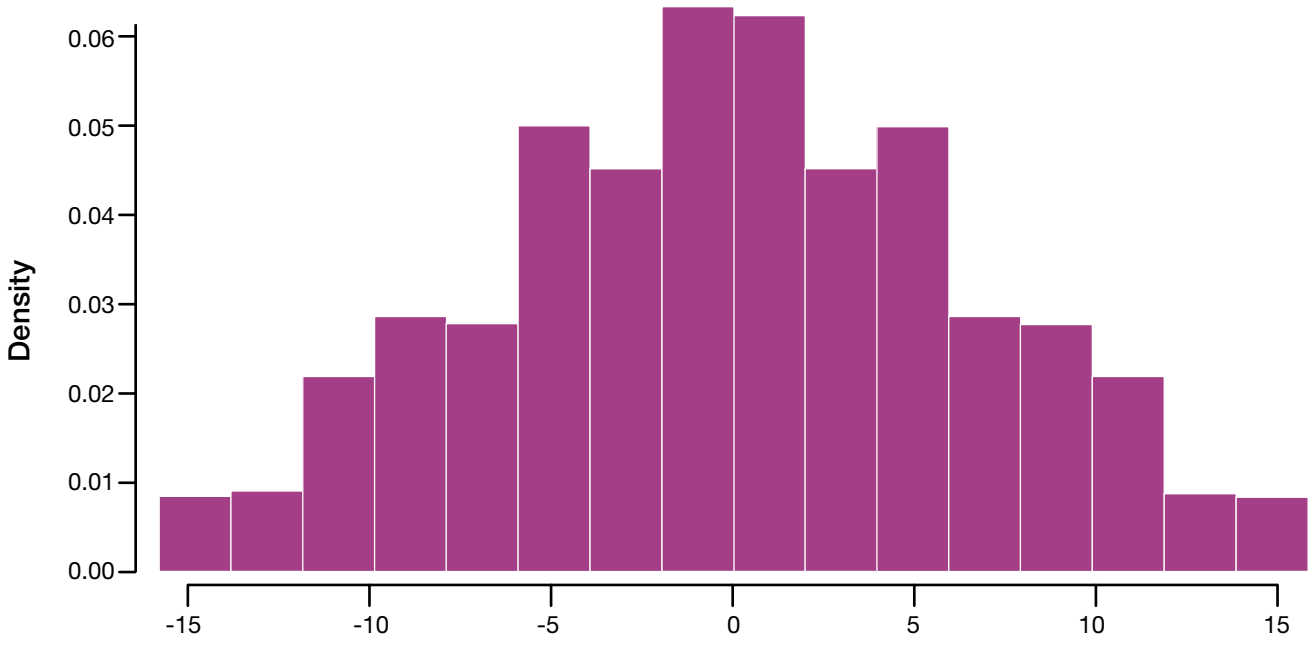
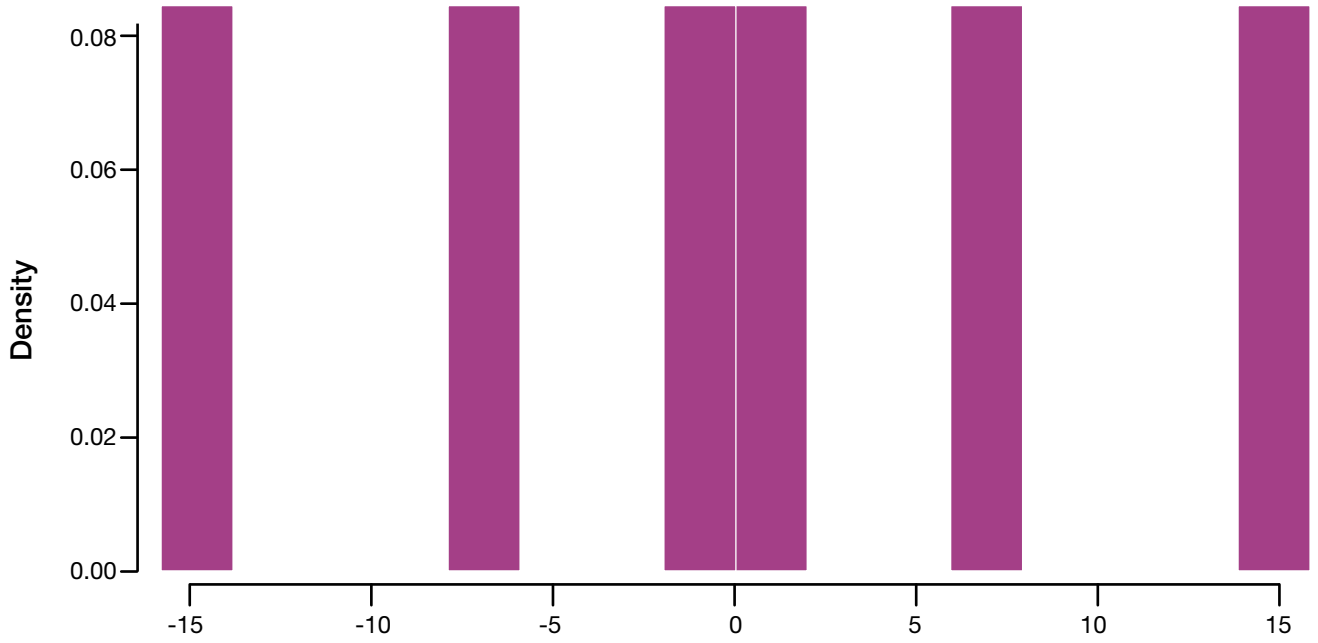


Figure 3. Histogram of Impacts When Families Are Randomly Assigned



When is it NOT necessary to adjust for a level of clustering in an RCT?

If a cluster is neither the unit of random assignment, nor the unit of random sampling, then there is no "need" to account for clustering. For example, in a study with three purposefully (not randomly) selected schools in which 50 students are randomly assigned to treatment and control groups within each of the schools (that is, assignment is blocked/stratified by school), there is no need to adjust for clustering of students within schools, because schools were not randomly assigned or sampled.

It is also unnecessary to adjust for intermediate levels of clustering if they are not randomly assigned or randomly sampled. For example, in a study in which schools are randomly assigned to treatment and control groups and which includes health classes from each school, there is no need to adjust for clustering of students within classrooms. The only clustering requiring adjustment is the clustering of students within schools. This adjustment is necessary because schools were randomly assigned, but classrooms were neither randomly assigned nor randomly sampled.

However, some researchers may choose to adjust for clustering even when they are not required to, if they deem it appropriate for the research question under consideration. One reason for adjusting for clustering when clusters were not randomly sampled or assigned is to generalize findings to some larger population (sometimes called a "super population").

What Can Be Done to Improve Statistical Power in a Clustered RCT?

Studies that randomly assign individuals can usually detect smaller impacts than studies that randomly assign clusters due to the larger variance of the impact estimate that results from clustering. That is, studies that assign clusters have less statistical power (Schochet 2008).

> Schochet, Peter. "Statistical Power for Random Assignment Evaluations of Education Programs." *Journal of Educational and Behavioral Statistics*, vol. 33, no. 1, 2008, pp. 62–87.

Reducing the variance of the impact estimate will improve statistical power in a clustered RCT. The variance of an impact estimate in a clustered RCT depends on primarily three factors: (1) the intraclass correlation (ICC), (2) the cluster-level regression R^2, and (3) the sample size (particularly the number of clusters).

The ICC is the proportion of total outcome variance that is due to between-cluster variance. A larger ICC leads to a greater effect of clustering. One strategy to reduce the ICC is to ensure that the study includes clusters (for example, schools) with outcomes of interest that are as similar as possible.

The cluster-level regression R^2 is the proportion of between-cluster variance that can be explained by covariates. A larger cluster-level regression R^2 reduces the effect of clustering. One strategy to increase the cluster-level regression R^2 is to collect data on baseline variables known to be highly correlated with the cluster-level mean outcome and then include those variables as covariates when calculating regression-adjusted impacts.

In all RCTs, a larger sample size will lead to a smaller variance. In a clustered RCT, increasing the sample size at the cluster level is usually more beneficial than increasing the sample size at the individual level. We illustrate this point in Table 1: Doubling the number of clusters reduces the minimum detectable impact (MDI) more than does doubling the number of individuals per cluster.

Table 1. Minimum Detectable Impacts Associated with Sample Size Increase at the Cluster and Individual Levels

	Number of Clusters	Number of Individuals Per Cluster	Total Number of Individuals	Minimum Detectable Impact (percentage points)
Original	**20**	**50**	**1,000**	**12.4**
Double clusters	40	50	2,000	8.7
Double individuals per cluster	20	100	2,000	11.2

Note: These calculations assume a dichotomous outcome with a prevalence rate of 30 percent, an ICC of 0.03, and no covariate adjustment.

How can you account for clustering when calculating the variance of an estimated impact in a clustered RCT?

Several statistical methodologies can be used to account for clustering when calculating the variance of an estimated impact in a clustered RCT. We suggest the following three methods:

1. ***Mixed Effects Modeling.*** Also known as "hierarchical linear modeling (HLM)" or "random effects," this approach to linear regression accounts for clustering using maximum likelihood to estimate parameters that specify the structure of the covariance between individuals in clusters.

   Hsiao, C. *Analysis of Panel Data, 2nd edition*. Cambridge, UK: Cambridge University Press (Econometric Society Monographs, No. 34), 2003.

   Raudenbush, S.W. and A.S. Bryk. *Hierarchical Linear Models (Second Edition)*. Thousand Oaks, CA: Sage Publications, 2002.

2. ***Generalized Estimating Equations (GEE).*** Also known as "the sandwich estimator," this approach to linear regression accounts for clustering without distributional or modeling assumptions (it is "nonparametric"). Rather, it accounts for clustering based on the average estimated covariance among observations within each cluster.

   Williams, R.L. "A Note on Robust Variance Estimation for Cluster-Correlated Data." *Biometrics*, vol. 56, 2000, pp. 645–646.

3. ***Between-Cluster Estimation.*** This simple approach uses ordinary least squares (OLS) regression on data that has been aggregated (in other words, "averaged," or "collapsed") to the cluster level.

   Hsiao, C. *Analysis of Panel Data, 2nd edition*. Cambridge, UK: Cambridge University Press (Econometric Society Monographs, No. 34), 2003.

What are the advantages and disadvantages of these approaches to accounting for clustering?

To assess the trade-offs in using the three methods described above, we conducted simulations of clustered RCTs. In our simulations, we created a dichotomous outcome with a prevalence rate of 20 percent, using different assumptions about sample size and covariate adjustment. We also conducted the same simulations using a continuous outcome. We describe these simulations in detail in the appendix to this FAQ. The following points highlight our most important conclusions:

A. ***All methods examined are always substantially better than no adjustment at all.*** This conclusion is not surprising—theory tells us that it is important to adjust for clustering, and the simulation findings are consistent with theory.

B. ***There are important differences between studies with a very small number of clusters and studies with a larger number of clusters.*** We simulated studies with 6 clusters, 30 clusters, and 150 clusters, all evenly divided between treatment and control groups. From the simulations with just 6 clusters, we found:

1. ***Covariate adjustment offers little to no benefit.*** For all the methods examined, there is little to no improvement in precision when adjusting for covariates. In fact for the "between-cluster estimation" method, adjusting for covariates actually *increases* the standard error. (When the number of clusters is 30 or 150, there is a precision gain from covariate adjustment).
2. ***All methods "work" when the clusters are equal.*** When the six clusters all include the same number of individuals, all three methods used to adjust for clustering yield accurate *p*-values.
3. ***GEE does not "work" when the clusters vary substantially in size.*** When the clusters vary in size, using GEE yields *p*-values that are too small—that is, GEE overstates statistical significance. This effect does not occur with other methods. (When the number of clusters is 30 or 150, GEE still overstates statistical significance but to a much smaller degree).

C. ***Between-Cluster Estimation Is Surprisingly Effective.*** In contexts other than clustered RCTs, this approach can lead to regression coefficient estimates with much larger variance than the other two approaches. This effect occurs because the approach does not take advantage of within-cluster variation in explanatory variables of interest (that is, variables on the "right-hand side" of the regression equation). But in a clustered RCT, treatment status does not vary within clusters (though other covariates might). Consequently, we see from the simulation results that this method

produces impact estimates that are just as precise as the other methods.[2] Due to the simplicity of this method, these findings suggest that it could be an appealing alternative to the more complex approaches, at least when the only regression coefficients of interest are those for variables that do not vary within clusters.

Can different cluster-adjustment methods lead to substantively different impact estimates?

If the true impact of the program is the same for all clusters, then the three methods should offer very similar impact estimates.

However, if program impacts vary across clusters, then different cluster-adjustment methods can yield different impact estimates (Schochet 2009). This effect occurs because the methods differ in how they weight clusters when calculating impacts. We recommend sensitivity analyses to assess whether conclusions regarding intervention effectiveness change under alternative methods.

Schochet, Peter Z. *The Estimation of Average Treatment Effects for Clustered RCTs of Education Interventions* (NCEE 2009-0061 rev.). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Endnotes

[1] The number of possible (unique) random assignments = $\frac{N!}{(N/2)!\bullet(N/2)!}$, where N is the number of units being assigned (assuming equal treatment and control groups). N! is the factorial of N. For example, 5! = 5*4*3*2*1 = 120.

[2] One case is a partial exception. When the number of clusters is small, covariate adjustment noticeably *increases* the standard error when using this method. But, when the number of clusters is small, the standard error when using this method *without* covariate adjustment is about the same as the other methods with or without covariate adjustment.

This update was written by John Deke from Mathematica Policy Research for the HHS Office of Adolescent Health under contract #GS-10F-0050L, Task Order No. HHSP233201300416G

TECHNICAL APPENDIX

In this technical appendix to the document, "Frequently Asked Questions About the Implications of Clustering in Clustered Randomized Controlled Trials (RCTs)," we describe the simulations we ran to answer the question, "What are the advantages and disadvantages of these approaches to accounting for clustering?" We also present tables of findings from those simulations.

Simulation Details

We used a Monte Carlo simulation to assess the performance of different methods to adjust for clustering in RCTs. For all simulations, we used the same basic procedure:

1. Randomly generate the following variables (all are normally distributed with mean 0 and variance 1):
   - **a.** *e*—an individual-level error term
   - **b.** *u*—a cluster-level error term
   - **c.** *X*—an individual-level covariate
   - **d.** *Z*—a cluster-level covariate
2. Construct a continuous variable, Y*, using the formula

$$Y^* = \sqrt{\rho} \cdot \left[\sqrt{R^2_{clu}} \cdot Z + \sqrt{1 - R^2_{clu}} \cdot u \right] + \sqrt{1 - \rho} \cdot \left[\sqrt{R^2_{ind}} \cdot X + \sqrt{1 - R^2_{ind}} \cdot e \right]$$

   with the following parameter values:
   - **a.** $\rho = 0.10$ [the intraclass correlation (ICC)]
   - **b.** $R^2_{ind} = 0.20$ [the individual-level regression R^2]
   - **c.** $R^2_{clu} = 0.50$ [the cluster-level regression R^2]
3. For simulations for which you desire a continuous outcome (*Y*), set $Y = Y^*$. For simulations for which you desire a binary outcome (*Y*) with prevalence rate *PR*, set *Y* equal to 1 if Y^* is less than the *PR*-th percentile of Y^* and 0 otherwise. For all of our simulations of binary outcome variables, we selected a prevalence rate of 20 percent.
4. Randomly assign half of the clusters to a treatment group and half to a control group. Note that Y is not a function of treatment status, meaning that the true impact of being assigned to the treatment group is zero.
5. Calculate impacts, standard errors, and *p*-values using the following four methods, each with and without covariate adjustment (for a total of eight impacts):
   - **a.** Naïve OLS (no adjustment for clustering)
   - **b.** Mixed Effects (also known as hierarchical linear modeling)
   - **c.** Generalized Estimating Equations (GEE)
   - **d.** Between Estimator (OLS regression using data aggregated to the cluster level)
6. Repeat steps one through five 10,000 times, saving the impact and standard error estimates from each replication.
7. Calculate the "true" standard error of the impact for each estimation method as the standard deviation in estimated impact across all 10,000 replications. Calculate the "true" type-one error rate as the proportion of times that the estimated *p*-value falls below 5 percent (this proportion should equal 5 percent if the estimation method is working properly).

We applied this procedure for various outcomes (binary and continuous) and sample sizes, and with or without covariate adjustment.

Simulation Findings

Results from the simulations described above are reported in Tables A.1 and A.2. The values reported in the table cells are the simulation-based estimates of the type-one error rate. As described above, this rate should equal 5 percent. Deviations from 5 percent indicate that a method is calculating standard errors inaccurately and/or unreliably.

Table A1. Simulation-Based Comparisons of Cluster-Adjustment Methods–Binary Outcome

	Naïve OLS		Mixed Effects		GEE		Between Estimator	
Covariate Adjusted?	Standard Error	Error Rate	Standard Error	Error Rate	Standard Error	Error Rate	Standard Error	Error Rate
6 schools: 20 students per school								
No	0.102	0.142	0.102	0.013	0.102	0.035	0.102	0.054
Yes	0.107	0.11	0.107	0.009	0.107	0.038	0.145	0.048
6 schools: 2 with 10 students, 2 with 50 students, 2 with 100 students								
No	0.1	0.357	0.095	0.037	0.1	0.098	0.093	0.057
Yes	0.097	0.233	0.096	0.023	0.097	0.082	0.125	0.049
30 schools: 10 with 10 students, 10 with 50 students, 10 with 100 students								
No	0.044	0.378	0.041	0.051	0.044	0.064	0.041	0.052
Yes	0.035	0.269	0.033	0.052	0.035	0.06	0.033	0.046
150 schools: 50 with 10 students, 50 with 50 students, 50 with 100 students								
No	0.02	0.371	0.019	0.052	0.02	0.057	0.019	0.058
Yes	0.015	0.28	0.015	0.049	0.015	0.051	0.015	0.047

Source: Monte Carlo simulations with 10,000 replications.

Table A2. Simulation-Based Comparisons of Cluster-Adjustment Methods–Continuous Outcome

	Naïve OLS		Mixed Effects		GEE		Between Estimator	
Covariate Adjusted?	Standard Error	Error Rate	Standard Error	Error Rate	Standard Error	Error Rate	Standard Error	Error Rate
6 schools: 20 students per school								
No	0.309	0.251	0.309	0.021	0.309	0.035	0.309	0.05
Yes	0.295	0.193	0.295	0.014	0.295	0.038	0.417	0.048
6 schools: 2 with 10 students, 2 with 50 students, 2 with 100 students								
No	0.335	0.498	0.303	0.04	0.335	0.101	0.306	0.056
Yes	0.307	0.386	0.288	0.044	0.307	0.091	0.392	0.054
30 schools: 10 with 10 students, 10 with 50 students, 10 with 100 students								
No	0.148	0.509	0.132	0.047	0.148	0.064	0.135	0.056
Yes	0.111	0.423	0.102	0.054	0.111	0.069	0.104	0.061
150 schools: 50 with 10 students, 50 with 50 students, 50 with 100 students								
No	0.067	0.512	0.059	0.051	0.067	0.056	0.061	0.065
Yes	0.048	0.426	0.044	0.052	0.048	0.052	0.045	0.057

Source: Monte Carlo simulations with 10,000 replications.