The effects of a dropout prevention program on secondary students' outcomes

Enrico Conti (Irpet)

Silvia Duranti (Irpet)

Alessandra Mattei (Università degli studi di Firenze)

Fabrizia Mealli (Università degli studi di Firenze)

Nicola Sciclone (Irpet)

PRELIMINARY VERSION

PLEASE DO NOT QUOTE OR CITE WITHOUT PERMISSION FROM THE AUTHORS

1. Introduction

The purpose of this paper is to examine and validate the effectiveness of INNOVARE, a teacher-based dropout prevention program, aimed at reducing the number of early school leavers through the introduction of innovative teaching methods in the early grades of vocational schools. Although many school-based prevention programs are on the rise, few have been empirically validated, mostly by program developers themselves (Blum & Ellen, 2002; Brooks, 2006, Cho, Hallfors, & Sanchez, 2005). Instead, effectiveness trials conducted by independent researchers are an ethical obligation and are also necessary for understanding dropout prevention programs in real-world settings.

The problem of high school dropout continue to afflict Italian youth, particularly in vocational school system, where a great part of students come from disadvantaged social classes and have a "poor" family background. Moreover, in these schools immigrant and disabled students are over-represented . In such an environment teaching is not an easy task and it is frequent for teachers to experience situations of "burn out". Indeed, it is also difficult to involve students in teaching programs and therefore to provide them with adequate basic skills.

2. The INNOVARE program

In the described context, the Innovare project is aimed at reducing early school leaving through the re–motivation of teachers and the innovation of teaching methods. To this end, the project applies the methods of social research called "Action Research" , which has proved successful in similar contexts (Kemmis & McTaggart, 1982). Teachers, properly guided by tutors who are disciplinary experts in education and epistemology, become the designers/creators of the new teaching method, in a process of continuous comparison - reflection - correction of the educational practices implemented. The new way of teaching conceives of the formalization and systematization of knowledge as a point of arrival of the process of teaching- learning. It starts with the identification of a concrete goal to reach (the design-production of a product or service, the solution of a specific problem) and is characterized by an extensive use of educational workshops. The innovative aspect of Innovare is that it starts from the training of teachers to motivate or re-motivate both teachers themselves and students.
The project activity consists of 10 meetings between the expert-tutors and the teachers involved in the project, which then lead to the application of the new proposed teaching to their students during the school year.
The project involves 18 first classes in 12 public vocational schools located in the Tuscan provinces of Florence, Pistoia, Lucca, Pisa and Massa Carrara.
The subjects considered by the experiment are: Italian, Mathematics, Foreign language, Integrated science, Physics and technology.
The evaluation design consists of two instruments:

- A quantitative statistical approach that verifies the presence or absence of a causal link between the treatment of the classes involved in the project INNOVARE and some outcome variables.
- A qualitative evaluation aimed at identifying strengths and weaknesses of the project through a Focus Group with stakeholders (teachers and tutors-experts).

3. Data

Information about students, teachers and classes involved in the treatment were collected from administrative sources and trough questionnaires addresses to both students and teachers. Total number of students involved in the program amounts to 429.

To conduct a proper evaluation of the project, the same information was also collected for 829 pupils in 35 classes so-called "control", not treated by experimental teaching but having some similar basic characteristics.

Administrative data, consisting on school registers, were provided by schools. Information on students' personal and family characteristics were collected though a questionnaire administered at the beginning of the second semester. A second questionnaire was administered to teachers in order to detect information on their age and contract type.

The information collected are summarised in Table 1.

Table 1. Collected Data

Administrative sources		**Individual questionnaire**	
Individual level information	**Class level information**	**To the teachers**	**To the students**
Hours of absence in the first semester	Class size at the beginning of the school year	Type of contract (fixed-term, open-ended)	Sex
Average mark in the first semester	Class size at the beginning of the second semester	Age	Year of birth
Hours of absence in the second semester	New entrants in the second semester		Nationality
Average mark in the second semester	% foreigners		Late/not late
Evaluation at the end of the year (admitted to the following class, failed, postponement of the evaluation)	% males		Level of motivation at the beginning of secondary school (high, low)
Drop-out	% late students		Parents' education level (primary education or higher)
	% repeating students		Parents' occupational status (employed, unemployed)
	% students with parents with a low education level		
	% students with unemployed parents		
	% low motivated students		
	% drop-outs in the first semester		

	% absence rate in the first semester		
	Average conduct mark in the first semester		
	Average mark in the first semester		

Notes. In the Italian schooling system, evaluation is given at the end of the first quarter (end of January) and at the end of the second quarter (which coincides with the end of the school year). At the end of the first quarter, only marks are given, while at the end of each year a student is given not only marks but also a synthetic evaluation which can be: admission to the following class, failure (and thus repetition of the same grade), postponement of the evaluation to September (when the student will be newly evaluated in view of the results of an exam).

4. The evaluation of the INNOVARE program

The INNOVARE study was a cluster-randomized trial promoted and conducted by the Tuscan Regional government (Italy) in collaboration with the Regional school administration, a Teachers' association (CIDI) and the Regional Institute for Economic Planning of Tuscany (IRPET). The study aimed at assessing whether innovative teaching methods (including lab sections) could reduce drop-out from high school of students in vocational schools.

In the INNOVARE study a random sample of 53 classes was drawn from the vocational schools participating in the study: 18 classes were assigned to an innovative teaching method (intervention group) and 35 classes were assigned to the control treatment. In control classes, standard lectures were provided. Therefore, in the INNOVARE study the unit of assignment is the class.

In a cluster-randomized trial (e.g., Donner, 1998; Murray, 1998; Braun and Feng, 2001; Frangakis, Rubin, and Zhou, 2002; and Murray et al., 2006), clusters are assigned to treatment or control, but often individuals are of interest. Thus the unit of assignment may be different from the unit of analysis. The choice of the unit of analysis, which mainly depends on the research question of interest, is crucial, because it drives the statistical procedures to apply to draw inference on the quantities (estimands) of interest.

In this paper, we conduct both cluster-level analyses and individual-level analyses using the potential outcome approach to causal inference (e.g., Rubin 1974, 1978, 1990a,b, 2005). A cluster-level analysis may provide useful information on the effectiveness of the intervention in reducing high-school drop-out. Here, drop-out is viewed as a social problem and focus is on interventions that can limit school drop-out as a whole. The class is the natural unit of inference and standard methods for the analyses of randomized experiments can be applied at the cluster level.

An individual-level analysis aims at assessing whether the innovative teaching method has a causal effect on the student probability of dropping-out of school. In this case, the unit of assignment (class) is different from the unit of analysis (student), and the lack of independence among student in the same class, i.e., the presence of intraclass correlation, creates special methodological challenge and cannot be ignored.

Both approaches have advantages and drawbacks. In a cluster-level approach, we can always conduct exact statistical inferences without introducing parametric assumptions and we can easily adjust for background characteristics. Also, a cluster-level analysis is correct and valid irrespective of the strength of the intraclass correlation, because it implicitly accounts for all sources of variability.

In the INNOVARE study, we implement cluster-level randomization inference adjusting for differences in the observed cluster-level covariates using subclassification on the propensity score (Rosenbaum and Rubin, 1983). The randomization inference is non-parametric in that it does not make any functional form assumption and it is exact in that it does not rely on large sample approximations. Thus, results coming out of this analysis are exact and valid irrespective of the number of group assigned to each treatment status (e.g., Small et al., 2008, Imbens and Woolbridge, 2009; Mealli et al., 2011).

An individual-level analysis which accounts for the presence of intraclass correlation, using e.g., mixed effect regression models, may lead to more powerful model-based tests if the model is well specified than group randomization inference (Braun and Feng, 2001). In our study, individual-level analyses are conducted using multilevel models. Exact statistical inferences are not routinely available for multilevel regression models, but statistical inferences are usually based on large sample theory (i.e., large numbers of clusters). In the literature, there exist widely accepted guidelines for the numbers of

cluster required to ensure validity of statistical inferences. Results obtained by model fitting using data from studies enrolling fewer than 20 clusters per intervention group should be interpreted with caution (Duncan et al., 1998). The INNOVARE study involves only 18 treated classes and 35 control classes. Therefore, statistical inferences drawing using student-level analyses based on multilevel models may result in not much powerful tests due to the small number of classes assigned to the new treatment.
When applicable, multilevel models offer several advantages over cluster-level analyses. Specifically, multilevel models allow one to (1) obtain estimates of intraclass correlation more naturally, which can be used to design future studies; (2) adjust for background covariates at both individual- and cluster-level; (3) investigate sources of heterogeneity in the treatment effect, including interaction between the treatment variable and some specific covariates; and (4) extend the analyses to more complex data structures more easily, involving more than two levels.

Let us now introduce some notation. There are K=53 clusters (classes). Each class contains n_k students, i=1, …, n_k; k=1,..., K. Therefore, there are $N = \sum_{k=1}^{K} n_k$ students in total. A fixed number of M=18 classes are randomly assigned to the active treatment and K-M classes are assigned to the control treatment. For each class k, let W_k denote the treatment assignment : W_k=0 for classes assigned to the standard treatment, W_k = 1 for those assigned the new treatment. Let $Y_k(w)$ be the potential outcome at cluster-level, given assignment to treatment level w. If the kth cluster is randomly assigned to treatment, write W_{ki} = 1 for all i=1, …, n_k; otherwise, if this cluster is assigned to control, write W_{ki} = 0 for all i=1, …, n_k. The ith student in the kth class has two potential responses, $Y_{ki}(0)$ if the cluster containing i is assigned to treatment, and $Y_{ki}(1)$ if this cluster is assigned to control. Let $Y_k = Y_k(W_k)$ and $Y_{ki} = Y_{ki}(W_{ki})$ be the actual outcome observed at cluster- and student level, respectively. In the study we focus on five outcomes: failure, postponement of the evaluation, drop-out, absence rate and the union of failure and drop-out. A vector of custer-level pre-treatment variables, $\mathbf{X}_k$, is observed for each class, and a vector of individual-level pre-treatment variables, $\mathbf{Z}_{ki}$, is observed for each student. Here, cluster-level pre-treatment variables include both class-specific characteristics and group-average of individual-level pre-treatment variables.

Table 2. Summary statistics of cluster-level variables

	Mean	
	Z_k=0	Z_k=1
Outcome variables		
Percentage of failures	22.503	19.541
Percentage of postponements of the evaluation	32.048	34.979
Percentage of drop-outs	9.276	9.061
Absence rate (%)	14.612	15.691
Percentage of failures + Drop-out	31.779	28.602
Pretreatment variables		
Number of enrolled students at the beginning of the academic year	23.686	23.833
Number of students still enrolled at the beginning of the second term	21.600	21.889
Number of students enrolled at the beginning of the second term	0.943	0.444

% of failures at the end of the first term	8.636	8.080
Absence rate at the end of the first term	11.492	12.787
Average behavior score at the end of the first term	7.642	7.303
Average score at the end of the first term	5.670	5.590
Percentage of foreign students	27.157	27.923
Percentage of male students	72.693	81.619
Percentage of delayed students	53.705	57.924
Percentage of remedial students	43.357	49.758
Percentage of students whose parents are low-educated	42.769	49.062
Percentage of students whose parents are unemployed	27.560	32.975
Percentage of low-motivated students	18.645	21.086
Teacher's position (tenured teacher versus fixed-term teacher)	0.857	0.889
Teacher's age	0.857	0.722

4.1 Cluster randomization inference

In this section, we focus on a cluster-level analysis, using classes as units of analysis. Therefore, only cluster-level variables enter the analysis. Randomization inference allows us to draw exact inferences using only the random assignment of clusters to treatment or control. In randomization inference focus is on the observed sample, therefore sampling issues do not matter. Also no assumption is made about the underlying model that generated the data and the dependence structure of random cluster effects. Finally, the issue of interference between students in the same cluster does not arise, because focus is on cluster-level. We can reasonably assume that students in different classes do not interfere with each other, therefore for each class there exist only two potential outcomes in this experiment: $Y_k(1)$ if cluster k is assigned to treatment, and $Y_k(0)$ if cluster k is assigned to control.

Table 1 shows summary statistics for the sample of 53 classes grouped by assignment, W_k. Although classes are randomly assigned to treatment, Table 1 shows that there exist some differences in background pretreatment variables between the treatment group and the control group. In order to account for these differences in the observed pretreatment variables we use subclassifications on propensity score - the conditional probability of receiving a treatment given pretreatment characteristics under the assumption that the treatment is strongly ignorable: $\Pr(W_k=1 \mid Y_k(0), Y_k(1), \mathbf{X}_k)=\Pr(W_k=1 \mid \mathbf{X}_k)$, and $0 < \Pr(W_k=1 \mid \mathbf{X}_k) < 1$, $k=1, \ldots, K$ (Rosenbaum and Rubin, 1983). Strong ignorability amounts to assuming that within cells defined by the values of pre-treatment variables, the treatment is randomly assigned. Under this assumption we can view INNOVARE as a stratified cluster randomized experiment (Small et al., 2008). Rosenbaum and Rubin (1983) show that if the exposure to treatment is random within cells defined by the covariates, it is also random within cells defined by propensity score: $\Pr(W_k=1 \mid Y_k(0), Y_k(1), e(\mathbf{X}_k))=\Pr(W_k=1 \mid e(\mathbf{X}_k))$, where $e(\mathbf{X}_k)=\Pr(W_k=1 \mid \mathbf{X}_k)$, is the propensity score for the *kth* class, $k=1,\ldots, K$.

In our analysis, the propensity score is estimated using a logit regression model. Based on the estimated propensity score, we restrict the analysis to the subsample of classes that satisfies an overlap or common support condition. Specifically, we discard four control classes with propensity score values lower than the minimum propensity score value for the treated

classes, and one treated class with propensity score greater than 0.9, which is an extremely high value in our sample. Then, we re-estimated the propensity score using the selected subsample of 48 classes and use it for adjusting treatment comparisons for differences in background covariates using subclassification.

We divided the sample into H=4 strata based on propensity score categories as shown in Table 2. Analyses aim at assessing the balancing property of the propensity score suggest that covariates are well-balanced between treated and control classes within propensity score strata.

The stratified cluster randomized experiment underlying INNOVARE can be described as follows. Let $\mathbf{W}=(W_1,...,W_K)'$. Let K_h denote the number of clusters in stratum h, and let M_h be the (fixed) number of classes assigned to treatment in stratum h, h=1,...,H. Let $B_K \in \{1,\ldots, H\}$ denote the stratum for class k. There are $L = \prod_{h=1}^{H} \binom{K_h}{M_h}$ possible values $\mathbf{w}=(w_1,...,w_K)'$ of the treatment assignment $\mathbf{W}$, and and each has probability 1/L. In our study, L= 120 · 1,287 · 36 · 45 = 250,192,800.

Under the assumption that data come from a stratified cluster randomized experiment, our cluster-level analysis use randomization inference to draw exact inferences for our finite population (sample) of size K=48. We adopt the Fisher Exact P-values approach (Fisher, 1925).

Table 3. Propensity score strata

Stratum	Propensity Score	Control Classes	Treated Classes	Total Number of Classes
1	0.00 – 0.20	14	2	16
2	0.20 – 0.40	8	5	13
3	0.40 – 0.51	7	2	9
4	0.51 – 1.00	2	8	10

Fisher focused on deriving exact p-values for sharp null hypotheses regarding the effect of treatments. Under a sharp null hypothesis all potential outcomes are known from the observed values of the potential outcomes. The most common null hypothesis in Fisher's framework is the s harp null hypothesis of no effect of the treatment for any unit (class) in the population: H_0: $Y_k(0) = Y_k(1)$ for all k, which implies that $Y_k(0) = Y_k(1) = Y_k$, for all k.

Under this type of null hypotheses, the value of any statistic S, that is, any function of the stochastic assignment vector, the observed potential outcomes, and the pretreatment variables, is known, not only for the observed assignment, but for all possible assignments. Thus, the distribution of any statistic generated by the randomization of the treatment assignment, can be deduced. This distribution is usually referred to as the randomization distribution. Using the randomization distribution of the statistic we can calculate p-values as the probability (under the assignment mechanism and under the null hypothesis) that we would observe a value of S as unusual as, or more unusual than, the observed value of S. Therefore the Fisher Exact P-values approach entails three steps: (i) the choice of a sharp null hypothesis, (ii) the choice of test statistic, and (iii) the measure of extremeness (p-values). Here we focus on the s harp null hypothesis of no effect of the treatment and we consider two test statistics: the difference in average outcomes by treatment status, S_{ave}, and the difference in average ranks for treated and control units , S_{rank}, which are defined as follows:

$$S_{ave} = \sum_h \frac{K_h}{K} S^h_{ave} = \sum_h \frac{K_h}{K} [\bar{Y}_{1h} - \bar{Y}_{0h}] = \sum_h \frac{K_h}{K} \left[\frac{1}{M_h} \sum_{k:B_k=h, W_k=1} Y_k - \frac{1}{K_h - M_h} \sum_{k:B_k=h, W_k=0} Y_k \right]$$

$$S_{rank} = \sum_h \frac{K_h}{K} S^h_{rank} = \sum_h \frac{K_h}{K} [\bar{R}_{1h} - \bar{R}_{0h}] = \sum_h \frac{K_h}{K} \left[\frac{1}{M_h} \sum_{k:B_k=h, W_k=1} R_k - \frac{1}{K_h - M_h} \sum_{k:B_k=h, W_k=0} R_k \right]$$

where R_k is the normalized rank :$R_k = R_k(Y_1, \ldots, Y_K) = \sum_{l=1}^{K} 1\{Y_l < Y_k\} + \frac{1}{2}\left(1 + \sum_{l=1}^{K} 1\{Y_l = Y_k\}\right) - \frac{K+1}{2}$.

The test statistics are calculated as weighted average of the test statistics across strata defined by the estimated propensity

score with weights given by the proportion of classes in each stratum. Table 3 shows the observed values of the test statistics and the *p*-values against the alternative that, at least for some units, there is a non-zero effect : H_1: ∃ *k*: $Y_k(0) \neq Y_k(1)$. The *p*-values are estimated using 10,000 draws from the randomization distribution. The test statistics show some evidence that the new teaching method reduces the percentage of drop-outs and failures and the absence rate, and increases the percentage of postponements of the evaluation.

Table 4. Observed values of the test statistics and p-values for the sharp null hypothesis H_0: $Y_k(0) = Y_k(1)$ ∀ *k* against the alternative H_1: ∃ *k*: $Y_k(0) \neq Y_k(1)$

Outcome variables	S_{ave}	***p*-value**	S_{rank}	***p*-value**
Percentage of failures	-2.78	0.6698	-1.78	0.7104
Percentage of postponements of the evaluation	5.87	0.2320	5.77	0.2270
Percentage of drop-outs	-2.41	0.7734	-3.12	0.7744
Absence rate (%)	-0.15	0.9434	-1.79	0.6804
Percentage of failures + Drop-out	-5.19	0.4554	-5.25	0.4794

It is worth noting that the observed values of the statistic S_{ave} are greater (in absolute term) than the differences in average outcomes by treatment status calculated without accounting for differences in background covariates (see Table 1). Therefore, adjusting for differences in background covariates emphasizes the effect of the new teaching method in the INNOVARE study. However the *p*-values do not show any evidence against the null hypothesis of no treatment effect.

4.2 Individual-Level Analyses based on Multilevel Models

In this section we propose an individual-level analysis based on multilevel models, which properly account for dependencies of responses for students from the same class and allow us to adjust for both individual-level and cluster-level characteristics. Recall that the number of classes assigned to the intervention group is relatively small in the INNOVARE study: only 18 classes are picked for the new teaching method. Therefore results shown in this section must be interpreted with caution.

We consider generalized linear mixed models with probit link for binary outcome variables and linear mixed models for continuous outcomes. Formally, let $\boldsymbol{C}_{ki}$ be the vector of all the explanatory variables (including the treatment indicator) for student *i* in class *k* included in a model. Then, $Y^*_{ki} = \beta_1 + \boldsymbol{\beta}_2 \boldsymbol{C}_{ki} + u_k + \varepsilon_{ki}$, where $Y^*_{ki} = Y_{ki}$ and $\varepsilon_{ki} \sim N(0,\sigma^2)$ if Y_{ki} is a continuous variable, and Y^*_{ki} is a latent variable such that $P(Y_{ki}=1) = P(Y^*_{ki} > 0)$ with $\varepsilon_{ki} \sim N(0,1)$ if Y_{ki} is a binary outcome. We also assume that $u_k \sim N(0,\sigma^2_u)$ independently of ε_{ki}.

For each outcome variable we consider two alternative model specifications, say A and B. Model A includes only individual-level and cluster-level characteristics as explanatory variables; Model B includes also group-averages of the first level variables as explanatory variables on top of individual-level and cluster-level characteristics. Group-averages of the first level variables allow us to account for the presence of correlation between individual-level variables and cluster effects, as well as for the presence of interference between students belonging to the same class.

The fitted models lead to very small (close to zero) estimates of the intraclass correlation coefficients, especially under Model B, suggesting that dependencies of responses for students from the same class tend to vanish

Table 5. Model-based estimates of the coefficients for the treatment variable, $W_{ki,}$ (standard errors in parenthesis), the group-level variances (residual variance), the average potential outcomes and the average treatment effect

Outcome variable	W_{ki}	**Variance:**	$E[Y_{ki}(0)]$	$E[Y_{ki}(1]$	$E[Y_{ki}(1)]-E[Y_{ki}(0)]$

		Cluster-level (σ^2_u) (Residual: σ^2)			
Model A					
Failures	-0.240 (0.137)	0.073	0.154	0.104	-0.050
Postponement of the evaluation	0.093 (0.010)	0.025	0.276	0.308	0.032
Drop-out	-0.131 (0.307)	0.628	0.012	0.008	-0.003
Absence rate (%)	-0.024 (1.044)	9.373 (52.824)	14.526	14.402	-0.124
Failure + Drop-out	-0.274 (0.156)	0.101	0.230	0.156	-0.074
Model B					
Failures	-0.258 (0.120)	0.000	0.157	0.105	-0.052
Postponement of the evaluation	0.245 (0.098)	0.000	0.253	0.337	0.084
Drop-out	-0.047 (0.183)	0.000	0.015	0.014	-0.002
Absence rate (%)	-1.024 (0.883)	4.284 (52.764)	14.835	13.810	-1.024
Failure + Drop-out	-0.217 (0.136)	0.000	0.223	0.164	-0.059

conditional on the covariates (see the second column in Table 4). Therefore the effective sample size, which is a decreasing function of the intraclass correlation, is high, close to the number of students in the sample, implying that the individual-level analysis may provide useful information.

The last three columns in Table 4 show the model-based estimates of the average potential outcomes and the average treatment effect (for binary outcomes these estimates are derived fixing the covariates at their observed mean). Our findings show some evidence that the new teaching method reduces the failure rate and the probability of either failing or dropping-

out. We also estimate that the new treatment reduces the absence rate, although the size of the effect is strongly influenced by the model specification. Finally, the new teaching method seems to increase the drop-out rate, even if the size of the effect is very small.

The coefficient for the treatment variable is never statistically significant under Model A, suggesting that the effect of the new teaching method is negligible. Conversely, Model B provides statistically significant evidence at the 5% level that the new teaching method reduces failure rate and increases the probability of postponement of the evaluation. The differences we observed between the alternative model specifications are at least partially due to extremely small intraclass correlations we estimate under Model B that includes group averages of the first level variables as explanatory variables. Generally speaking a small intraclass correlation implies a high effective sample size, which in turn leads to smaller standard errors.

REFERENCES

Braun, T.M., and Feng, Z. (2001) Optimal permutation tests for the analysis of group randomized trials. Journal of the American Statistical Association, 96, 1424-1432.

Donner, A. (1998) Some aspects of the design and analysis of cluster randomized trials. Applied Statistics, 47, 95-113.

Duncan C., Jones K., Moon G. (1998) Context, composition and heterogeneity: using multilevel models in health research. Social Science & Medicine 46, 96-117 .

Fisher, R.A. (1925) Statistical methods for research workers. Oliver and Boyd, Edimburgh, First edition.

Frangakis, C.E., Rubin, D.B., and Zhou, X.H. (2002) Clustered encouragement designs with individual level noncompliance. Biostatistics, 3, 147-177.

Imbens, G.W., and Wooldridge J.M. (2009) Recent developments in the econometrics of program evaluation. Journal of Economic Literature, 47, 5-86.

Mealli, F., Pacini, B., and Rubin, D.B. (2011) Statistical inference for causal effects. In Kenett R. and Salini S. (Eds.) Modern Analysis of Customer Satisfaction Surveys, Wiley, 173-192.

Murray, D. (1998) Design and analysis of group randomized trials, New York: Oxford University Press.

Murray, D., Hannan, P.J., Pals, S.P., McCowen, R.G., Baker, W.L., and Blitstein, J.L. (2006) A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. Statistics in Medicine, 25, 375-388.

Rosembaum, P.R., and Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41-55.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66, 688-701.

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

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

Rubin, D.B. (1990b). Formal modes of statistical inference for causal effects. Journal of Statistical Planning and Inference, 25, 279-292.

Rubin, D.B. (2005) Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association, 100, 322-331.

Small, D.S., Ten-Have, T.R., and Rosenbaum P.R. (2008) Randomization inference in a group-randomized trial of treatments for depression: Covariate adjustment, noncompliance, and quantile effects. Journal of the American Statistical Association, 103, 271-279.