The Canadian Journal of Program Evaluation Vol. 24 No. 2 Pages 31–56
ISSN 0834-1516 Copyright © 2010 Canadian Evaluation Society

A PROMISING NEW APPROACH TO ELIMINATING SELECTION BIAS

Laura R. Peck
School of Public Affairs, Arizona State University
Phoenix, Arizona, USA

Furio Camillo
Ida D'Attoma
Dipartimento di scienze statistiche, Università di Bologna
Bologna, Italy

Abstract: This article presents a creative and practical process for dealing with the problem of selection bias. Taking an algorithmic approach and capitalizing on the known treatment-associated variance in the X matrix, we propose a data transformation that allows estimating unbiased treatment effects. The approach does not call for modelling the data, based on underlying theories or assumptions about the selection process, but instead calls for using the existing variability within the data and letting the data speak. We illustrate with an application of the method to Italian Job Centres.

Résumé : Cet article décrit un processus inventif et pratique pour remédier au problème du biais dans la sélection. En se servant d'une approche algorithmique utilisant la variation connue associée au traitement dans la matrice X, nous proposons une transformation des données qui permet l'estimation des effets de traitement sans biais. Cette approche ne demande pas la modélisation des données basée sur des théories ou suppositions sous-jacentes quant au procédé de sélection. Notre proposition préconise plutôt l'utilisation de la variation existante dans les données et laisse parler les données mêmes. La méthode est illustrée en l'appliquant aux Centres d'Emploi Italiens.

INTRODUCTION

Selection bias may be the most important vexing problem in program evaluation or in any line of research that attempts to assert causality. In fact, the world's most prestigious award—the Nobel Prize—honoured James J. Heckman in Economics in 2000 "for his

Corresponding author: Laura R. Peck, ASU School of Public Affairs, 411 North Central Avenue, Suite 400, Phoenix, AZ, USA 85004-0697; <Laura.Peck@asu.edu>

development of theory and methods for analyzing selective samples" (Kungl Vetenskapakademien, 2000). Some of the greatest minds in economics and statistics have scrutinized the problem of self-selection, with the resulting approaches—Rubin's Potential Outcome Approach or Heckman's structural equation approach—being widely accepted and used as the best fixes. That said, these solutions to the bias that arises from self- or administrator-selection are imperfect, and many researchers reserve their strongest causal inference for data from experimental rather than observational studies.

In the conventional program evaluation context, problems associated with selection bias arise when one aims to assess an intervention's effectiveness but must rely on poorly matched comparison groups. For example, people who volunteer to participate in a program are different from those who do not volunteer in ways that affect their outcomes. Commonly, the classically designed experiment is considered an optimal solution; but for many reasons it may not be feasible to use an experimental evaluation design. In response, a variety of other non-experimental or quasi-experimental evaluation designs are used in attempts to assess program effectiveness, reflexive and non-equivalent group comparisons being common. Various approaches to imposing statistical controls also exist, one of which is the use of a propensity score to generate matched groups.

In Rubin's Potential Outcome Approach, the aim of the resulting "propensity score" is to balance non-equivalent groups on observed pre-treatment covariates in order to reduce bias in causal effect estimation. Rubin demonstrated that, having pre-treatment information that characterizes the units under analysis, it is possible to create groups of units with similar pre-treatment characteristics. These groups are, therefore, theoretically independent from the treatment. Within these groups, one then compares the target variable among those who have undergone the treatment and those who have not. The success of this propensity score matching approach in reducing bias mainly depends on the balance criteria adopted. It has recently been a hot topic in program evaluation (e.g., Thoemmes, 2009), with the verdict still out regarding its utility (e.g., Agodini & Dynarski, 2001; Dehejia & Wahba, 1999, 2002; Luellen, Shadish, & Clark, 2005; Wilde & Hollister, 2002).

To measure data imbalance, we follow the idea introduced in D'Attoma (2009) and Camillo and D'Attoma (2010) of summarizing the multivariate difference in covariates across treatment groups in terms of between-group inertia and then testing it with a multivariate imbalance test. "Inertia" is a measure of association among categorical covariates.[1] Grounded in a non-theoretical, statistical approach—rather than in economic theory, which requires some subjectivity—this approach tackles the problem of selection bias within the framework of multivariate descriptive analysis using patterns in the data to measure variation associated with treatment status. Specifically the approach involves the use of Conditional Multiple Correspondence Analysis (Escofier, 1988) as a tool for investigating the dependence relationship between covariates and the assignment-to-treatment indicator variable within a strategy whose final aim is to find balanced groups (D'Attoma, 2009).

Following recent work, we present this approach for estimating unbiased treatment effects, which combines a global imbalance (GI) measure and a multivariate imbalance test (Camillo & D'Attoma, 2010; D'Attoma, 2009). In brief, the method involves computing and testing the global imbalance, classifying cases in order to generate well-matched comparison groups, and then computing the treatment effect. The remainder of this article summarizes this approach, first by discussing the underlying paradigm and by explaining the problem of causal inference, then by describing the proposed method, and finally by illustrating an application.

UNDERLYING PARADIGM

Although others most certainly provide both a more thorough and a more nuanced discussion of the differences between the economic and statistical approaches, our attempt here is to make some observations about the two paradigms, and to describe the paradigm that underlies our proposed approach to dealing with selection bias. Generally, the economic approach is one that rests on underlying economic theory to drive and test models of economic behaviour and phenomena. Dealing with issues of selection bias in a program evaluation setting generally means modelling the selection process as a function of known variables. The persistent imperfection is the omnipresence of "unobservables" that one hopes are sufficiently dealt with by controlling for observables. Researchers acknowledge these shortcomings in their analyses and explore the implications of unobserved factors on the extent and direction of bias in results.

In contrast, a focus of statistics may be to fit the "best" model, but that model need not necessarily be based on some underlying theory about human behaviour. According to Breiman (2001), 98% of statisticians engage in a "data modelling culture" that emphasizes model validation through goodness-of-fit tests and residual examination; the other 2% represent an "algorithmic modelling culture," where predictive accuracy validates models (p. 199).

The technique examined in this article comes from the edge of the statistical perspective—Breiman's (2001) less common paradigm—where a fundamental underlying belief is that any researcher influence unduly affects the results, such that multiple "solutions" arrive simply by virtue of researchers' choice of model. More precisely, the paradigm we refer to is not only about statistics and economics, but about Breiman's (2001) two different cultures in statistical modelling: data modelling versus algorithmic modelling. The latter of these, the "data mining" perspective, can be thought of as "letting the data speak." This line of research compels questions about what the model is for a data miner, if the model suits the nature of the data, and if the model can correctly represent the real complexity of the data.

Breiman's (2001) work is fundamental to understanding the role and the limitations of data models and the rationale for utilizing, and perhaps even preferring, algorithmic models. He asserts that "[a]pproaching problems by looking for a data model imposes an a priori straight jacket that restricts the ability of statisticians to deal with a wide range of statistical problems" (p. 204). Conclusions from the data modelling perspective "are about the model's mechanism, and not about nature's mechanism," such that if a data model "is a poor emulation of nature, the conclusion may be wrong" (p. 202). If different models "give different pictures of the relation between the predictor and response variables," then "the question of which one most accurately reflects the data is difficult to resolve" and does not help in supporting useful conclusions for commercial or policy purposes (p. 203).

In brief, our underlying paradigm is that the problem at hand should define the solution's approach. In response to this observation, we propose to follow an algorithmic approach as appropriate to resolve the particular problem of bias in treatment effect estimation. As Camillo and D'Attoma (2010) point out—with reference to conventional propensity score methods—the subjectivity present in choosing which variables to use and testing balance and stratifying scores in propensity score estimation introduces bias into an analysis, which could much preferably be conducted via an algorithmic modelling approach. That said, the literature on propensity score estimation lacks objective criteria for selecting the best model, or for making other objective choices in the analytic process. The results are a "multiplicity of good models" with potential for informing wrong decisions (Breiman, 2001, p. 206). The choice of the best model should be justified according to a rigid, unbreakable criterion, quantitatively defined prior to the analysis (Camillo & D'Attoma, 2010). We suggest that our approach reduces bias in subsequent treatment effect estimation and leads to an optimal solution, without researchers' subjective decisions regarding functional form and estimation methods.

THE PROBLEM OF CAUSAL INFERENCE

The main problem of causal inference is essentially one of missing data. That is, in order to know whether some variable causes change in another variable for some unit, it is necessary to observe that unit in both its treated and untreated states, all else being equal. This observation is never possible, though in practical applications we estimate it all the time. Impact evaluation aims to measure the effect of some intervention on an outcome of interest among a population or subgroups. To do so, we must know the outcome that would be observed in the absence of the intervention as well as in the presence of the intervention, both measured at the same point in time (Rubin & Waterman, 2006). In other words, the missing element is the counterfactual outcome, defined as what would have happened in the absence of the intervention.

Researchers have taken various approaches to resolving the missing counterfactual problem. One common approach to establishing causality, particularly popular in economics, is the Structural Equation Model. This approach involves the specification of a model in which the response variables ("endogenous") are represented as functions of other, both exogenous and endogenous, variables and of external "error" variables. This approach differs from our point of view because it requires assumptions about the joint distribution of error terms, and it requires (subjective) economic theory underlying the model. Hence, the choice of right variables and the omitted variable become fundamental. Heckman, Ichimura, and Todd (1997) show that omitting important variables can seriously increase bias in resulting impact estimates.

Another approach is that of Pearl's (2000) graphical theory of causality. Pearl has shown how graphs provide a different way of thinking about causal relationships between variables and identification strategies that can be used to estimate causal effects. From the algorithmic modelling culture's perspective, we underline one important characteristic of Pearl's approach: the framework is completely nonparametric, and, as a result, it does not require specifying which variables have a causal relationship with the outcome.

Yet another common approach to the counterfactual problem, popular in a variety of areas (e.g., economics, medical research, epidemiology, and education), is that of the Potential Outcome Framework, pioneered principally by Rubin (1974, 1978). This approach is the closest to the algorithmic (or data mining) point of view. The main idea is that each individual in a population of interest has a potential outcome under each treatment state, even though each individual can be observed in only one treatment state at any point in time. With the potential outcome of interest for the treated as Y(1), the potential outcome for the untreated as Y(0), and the treatment variable, T (1 for treated, and 0 for untreated), we can observe combinations of {T=1, Y(1)} and {T=0, Y(0)} while we do not observe combinations of {T=0,Y(1)} and {T=1, Y(0)}, as depicted in Figure 1.

Figure 1
Observed and Missing Elements in the Potential Outcome Framework

Outcome Treatment	Outcome for Unit Exposed to Treatment, Y(1)	Outcome for Unit not Exposed to Treatment, Y(0)
Exposed to treatment, T=1	OBSERVED	MISSING
Not exposed to treatment, T=0	MISSING	OBSERVED

Given that we cannot observe the same unit under both treatment and control states, it becomes impossible to observe directly the causal effect of the treatment T on a specific unit i. Researchers then focus their attention on estimating the average causal effect, which is made particularly problematic when the assignment-to-treatment mechanism is not random and each potential outcome could belong to a different population.

The Potential Outcome Framework overcomes this problem by taking into account the selection mechanism. By modelling the selection mechanism using the propensity score model (Rosenbaum & Rubin, 1983), one can estimate the average treatment effect (ATE) in the context of non-experimental studies as $E[Y_i(1)]—E[Y_i(0)]$. In fact, the propensity score, defined as the probability that a unit gets assigned to a specific treatment given baseline covariates $[Pr(T_i=1 | X_i)]$, allows researchers to approximate a randomized experiment by eliminating part of the bias due to the selection mechanism. With the estimated propensity score, units under different treatment conditions but with similar values of propensity score, as predicted by their baseline covariates, can be compared. Unfortunately, in the evaluation context the rule for the assignment mechanism is often unknown and, as a consequence, the propensity score is not known; thus researchers must estimate it. In essence, in our view, *estimation* is the main problem.

A STRATEGY TO ELIMINATE SELECTION BIAS

The strategy proposed here moves from the concept that two units could be similar with respect to an objective distance measure. A pioneering work in this direction is Peck's (2005) "Using Cluster Analysis in Program Evaluation," where clustering algorithms are used to detect treatment group heterogeneity and create subgroups and subsequently measure causal treatment effects. Although that work relies on the presence of a random assignment mechanism, that algorithmic approach motivates the technique proposed in Camillo and D'Attoma (2010) and D'Attoma (2009) and summarized here with some extensions.

The innovative aspect of our proposed algorithmic approach is its reduction of selection bias under non-experimental conditions. Although both the structural equation model (Heckman) and the potential outcome framework (Rubin) aim to resolve the selection bias problem in causal inference, here we use Rubin's framework as our springboard. Specifically, our interest in the potential outcome framework arises from two important aspects. First, what Rubin (2006) calls "The Science" is represented by in our approach the information matrix (X), the observed potential outcomes (Y_obs) and the assignment-to-treatment indicator vector (T).

Second, as in the potential outcome framework, it is not necessary to collect the outcome variable in the design stage. Rubin (2007) asserts that "no outcome data from the study are in sight when objectively designing either a randomized experiment or an observational study" (p. 25). In practice, this analysis involves testing whether X and T are unrelated by subgroups ("clusters" in our strategy; "propensity score strata" in Rubin's strategy), then collecting outcome for units being compared. Just as Rubin does not consider the outcome variable in the propensity score model, we do not consider the outcome variable in the conditioning process.

In our approach, given an information matrix **X**, the analysis detects the variability associated with the selection mechanism (measured by between-group inertia) and creates a new space that is void of any variability associated with selection. Such an analysis stems from the observation that a matrix's overall variability [Inertia ($X;T$)] can be decomposed into elements that are independent of the selection mechanism [Inertia ($X \perp T$)] and dependent on that mechanism [Inertia ($X|T$)]. According to a conventional data matrix decomposition into eigenvalues and eigenvectors, the approach involves decomposing the portion of that matrix that does not depend on the selection mechanism. Applying such logic, we quantify the amount of conditioning (or "imbalance in data" or "between-group inertia") and thereby eliminate selection bias by estimating treatment effects only if conditioning results are insignificant. As a result, comparing treated and untreated units generates an unbiased impact estimate.

Combining these concepts, we provide evaluators with a simple three-step approach for estimating unbiased treatment effects in non-experimental data:

1. Measure and test balance: compute GI measure on the whole sample, and perform hypothesis test to evaluate its significance.

2. If imbalance exists, perform a Subgroups Analysis that involves the following substeps:

    a. Use Multiple Correspondence Analysis (MCA) to obtain a low-dimensional representation of the X-space.

    b. Apply Cluster Analysis to identify homogeneous groups on the basis of the low-dimensional MCA coordinates.

3. Measure and test balance within each group, and compute local ATE within balanced groups, pruning observations in unbalanced clusters.

This approach is a sort of subgroup analysis whose primary concept consists in getting finer cluster partition because it enhances the plausibility of obtaining balanced groups, thereby minimizing selection bias. The Appendix details the technical aspects of our approach that some readers might like to know, and the rest of this section elaborates on each of the three steps for the general evaluation practitioner.

Step 1: Measure and Test Imbalance

The first step involves measuring imbalance via the GI measure and then testing the extent to which there is imbalance in the data. In other words: Are there differences between treatment and comparison groups such that a simple comparison of their outcomes may be biased by selection? Such difference is measured in terms of between-group inertia, which represents a global measure of imbalance in data. The advantage of this GI measure stems from the consideration that most common variable-by-variable imbalance measures, such as difference in means or in proportions between treatment groups, might fail to detect imbalance. For example, if 4 of 10 baseline covariates are statistically different between treated and untreated cases, then conclusions about the presence or absence of imbalance are indecisive. In addition, variable-by-variable measures do not take into account any interactions between or among variables

Step 2: Cluster Analysis on MCA Coordinates

Within data that demonstrate the presence of selection bias, we proceed to the second analytic step, which involves executing a Cluster Analysis that identifies homogeneous groups on the basis of the low-dimensional MCA coordinates. This is known as the "Tandem Approach" (Arabie & Hubert, 1994). Using MCA coordinates before clustering exploits the advantage of working with continuous variables (MCA coordinates) rather than categorical covariates (original variables), which need to be treated with unusual metrics.

Step 3: Estimate Impacts

Next we assess the balance within Step 2's resulting clusters, computing local ATE within balanced groups (and pruning observations in unbalanced clusters). To a certain extent this step mirrors that of the step in a propensity score analysis where one identifies treatment and comparison group cases that are matched according to their assigned propensity score, and using those T-C pairs (or groups) to estimate a treatment effect. In this case, we use the outcomes from resulting treatment and comparison cases that are assigned in each cluster to generate the treatment impact for that subgroup.

APPLICATION TO ITALIAN EMPLOYMENT ASSISTANCE

In the area of employment training, selection bias has been shown to radically affect estimates of program impacts (e.g., Barnow, 1987;

Bloom, 2005, p. 182). In his review of the U.S. Comprehensive Employment and Training Act (CETA) programs from the 1970s, for example, Barnow (1987) shows that answers about program effectiveness depend on the measures and methods used. The illustration presented here, therefore, is useful since it applies our proposed analysis process to data from the Workfare Action program of the Italian *Centri Per l'Impiego* (CPI, hereafter also referred to as "Job Centres").

In the province of Bologna, as in the rest of Italy, the past two decades' labor market shifts have resulted in fewer "traditional" jobs (full-time, stable jobs with benefits) and more part-time or temporary jobs (OECD, 2007). This has placed those already vulnerable in the labor market (youth, immigrants, women, and early school leavers) in an even more precarious position. The Italian Job Centres aim to assist in the employability of these disadvantaged workers within the context of worsening labor market conditions. Understanding the effects of these government efforts is important but hindered by the observations that (a) they are targeted at a population whose characteristics predict worse employment outcomes, and (b) those who avail themselves of CPI services differ from those who do not in ways that also affect their employment outcomes. This situation poses a classic selection bias challenge. One source of selection bias suggests better outcomes and the other source suggests worse outcomes. A comparison of people who participate in the Job Centres' programs with those who do not participate would lead to spurious conclusions about program effectiveness. Further, in the Italian case, we do not have the option to randomly assign those who volunteer to participate to treatment and control status, and so we are left to deal with selection bias retrospectively and statistically, rather than prospectively through evaluation design.

One of the principal aims of the Job Centres is to reinforce the level of equality in a strongly differentiated workforce, by trying to compensate for the disadvantaged status of the "weaker" segments of the workforce in terms of job opportunities and the quality of the work that they are eventually offered. The "negative" differential characteristics of those people registered with the Job Centres are reflected in their employment outcomes.

The data come from the Province of Bologna's *Sistema Informativo Lavoro Emilia Romagna* (SILER; Work Information System) database. This archive includes 44,175 observations of job start-ups of short-term contracts between 2006 and 2008. Of these, 11,523 are individuals affiliated with the Bologna Job Centre and 32,652 are of individuals not affiliated in the Province.

To offer a sense of our database's characteristics, Table 1 shows selected baseline traits separately for people starting jobs who are affiliated with the Job Centres (our "treatment" group), and for those starting jobs who have never been affiliated with a Job Centre (the "comparison" group). It is clear that dependence between treatment status and each baseline covariate exists, since across all traits the treatment and comparison groups differ, and in some instances quite substantially. For instance, a greater proportion of those who seek Job Centre services are male (63.0% versus 49.7% in the comparison group), are in their thirties (33.0% versus 25.9% in the comparison group), and are known to be single or divorced (76.9% versus 45.6% in the comparison group), for example. We provide this descriptive profile only to make the point that the treatment and comparison groups differ from one another.

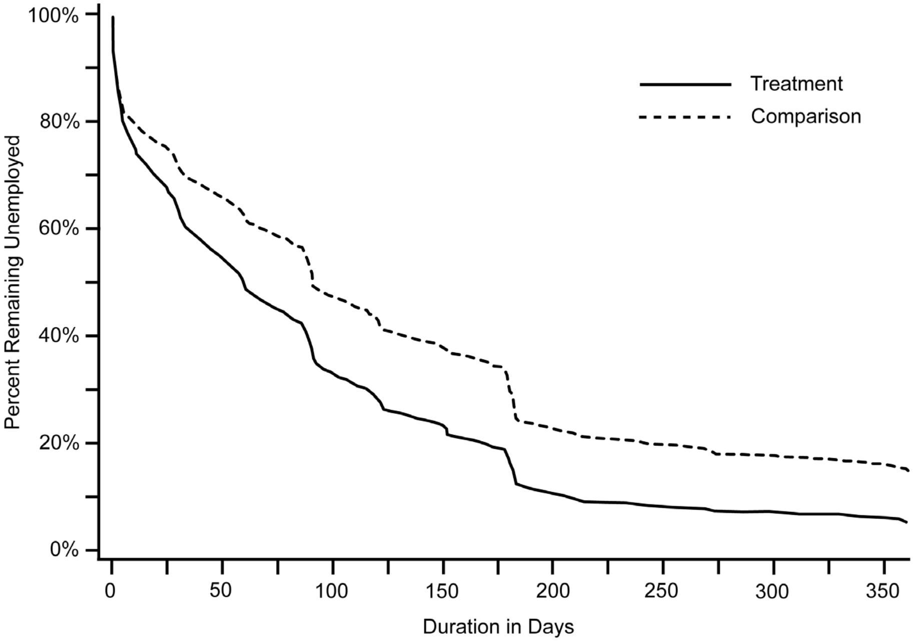
Table 1
Selected Baseline Characteristics, by Treatment Status

Baseline variable	Chi-Square	Overall	Treatment	Comparison	Difference
Sex (% Male)	42.5**	46.8	63.0	49.7	13.3
Age group (%): <2 5	35.7**	17.4	11.1	18.6	-7.5
25-30		26.2	25.8	26.4	-0.6
31-40		27.8	33.0	25.9	7.1
> 40		28.4	29.7	28.0	1.7
not available		0.2	0.1	0.2	-0.1
Marital Status (%): single	331.8**	21.0	35.1	16.1	19.0
married		0.8	1.3	0.6	0.7
divorced		32.7	41.8	29.5	12.3
separated		0.8	1.9	0.4	1.5
widowed		1.8	3.1	1.4	1.7
live-in-partner		0.2	0.3	0.1	0.2
not available		42.7	16.3	52.1	-35.8
Area of birth (% Bologna)	43.1**	35.0	30.2	36.6	-6.4
Area of residence (% Bologna)	302.9**	66.1	86.9	58.8	28.1
EU Citizenship (% member)	4.1*	14.0	16.1	13.2	2.9

Notes. *Treatment-comparison difference statistically significant at the $p < 0.05$ level. **Treatment-comparison difference statistically significant at the $p < 0.01$ level.

Assessing a causal link between the intervention and its employment outcomes is not easy, and it requires the effects of the Job Centres' actions to be disentangled from the influences of the population's characteristics. We expect that these differences in characteristics might explain differences in employment outcomes that are distinct from the contributions of employment assistance, although the expected direction of the bias is not obvious. Figure 2, calculated using the methods adopted by Cox and Oakes (1984), reports the survival curves of job start-ups—or, conversely, the end of unemployment—by treatment group status, and shows that the probability that the length of unemployment is consistently (and statistically significantly) lower among those affiliated with the Job Centres than it is among those in the comparison group. Although the lines are the same over the first days of follow-up, and the first 20% of employment happens at the same rate, thereafter, those registered with the CPI show quicker rates of terminating their unemployment. It is not clear, at this point, whether this difference is due to the program, to selection bias, or to any other source of bias, but what is clear, by our tests of statistical significance, is that the trends between the treatment and comparison groups are different.

Figure 2
Survival Curves of Job Start-ups, by Treatment Status



Note. We ran two tests equality of the survivor functions between groups: the Log-rank (chi-square=122.08; $p < 0.01$) and the Wilcoxon (chi-square 109.67; $p < 0.01$). These both support the conclusion that the curves are different from one another.

As we stated earlier, two sources of selection bias are likely at play: selection on observables may suggest less favourable employment outcomes, but selection on unobservables suggests relatively more favourable employment outcomes. We cannot know which of these forces is stronger, but ultimately that does not matter; what matters is eliminating that bias from the data so that we can estimate program impacts without the influence of selection bias. Our analysis considers all available data and does not judge the relationship between selection on these particular reported traits and treatment status; instead, we recognize that it exists and algorithmically adjust for it.

With this as context, we begin implementing the three-step analysis by computing the GI measure for this data set. As reported in Table 2, the resulting value of 0.042 can be interpreted as demonstrating the presence of imbalance in data. The GI measure falls in the critical region, thereby demanding adjustment in order to estimate a treatment effect that is not biased by selection.

Table 2
Balance in the Overall Sample

Treatment	Comparison	GI	Interval	Balance
11,523 26.1%	32,652 73.9%	0.042	(0;0.0033)	no

The second step in our analytic process is to use Cluster Analysis to identify homogeneous groups on the basis of the MCA coordinates. MCA was carried out using the following variables: gender (male/female), age group (<25; 25–30; 31–40; >40), marital status (single/married/divorced/separated/widowed/live-in partner/missing), area of birth (Bologna area or otherwise), nationality (EU member or otherwise), area of residence (Bologna area or otherwise), area of home address (Bologna area or otherwise), citizenship (EU or otherwise). The result of the MCA is a set of new variables (factorial coordinates) that are continuous and orthogonal to one another.

On the basis of these new MCA coordinates, we perform a cluster analysis to generate homogeneous groups of job start-ups during the 2006 through 2008 period. The cluster analysis was carried out on the SAS system employing Ward's algorithm on the MCA coordinates where the proximity between two groups is taken to be the square of the Euclidean distance between them. We most closely examined 60 cluster solutions. We retain the 30-cluster solution because it provides balance within a suitable number of clusters with fewer pruned observations (around 20%), compared to larger cluster solutions. The results reported in Table 3 represent a solution that represents a good trade-off between the number of clusters and the number of pruned observations. With the 30-cluster solution, we test the balance within each group, again using our computation of the GI and whether it falls in the critical region, as described in the prior step.

Table 3 shows the results of this cluster analysis in terms of balance, including the proportion of treatment and comparison cases that each cluster includes. Additionally, it includes *p*-values of the estimated difference between the treatment and comparison groups' survival curves. For data confidentiality reasons we cannot report characteristics of those cases within each cluster, but, given the atheoretical underpinning of our approach, we assert that each cluster's characteristics are irrelevant. That said, we recognize that in other practical applications, the analyst may be interested in characterizing the clusters. In this illustration, six of the clusters include either only treatment cases or only comparison cases, meaning that there is no common support for the subsequent impact estimation. Further, two of the clusters result in having unbalanced characteristics by our GI measure. In total, these eight clusters represent about 20% of the observations of the original sample being excluded from our third analytic step.

During the final stage of the procedure, we compare the job start-ups (or the termination of unemployment) for those workers affiliated at Job Centres with those not affiliated, separately within each of the remaining 22 clusters. We measure the effect of CPI affiliation on the length of unemployment. This choice of outcome measure is of interest to policy makers as they monitor the effect of CPI policies. Measuring unemployment duration depends on the passing of time, and varies according to the temporal distance from a job start-up. It also involves data truncation—that is, some terminations have not yet occurred and may not for a long time. It is appropriate, then, to use duration analysis, or survival analysis (Cox & Oakes, 1984), to consider the effects that CPI services are having on employment outcomes.

In particular, two survival curves (one for Job Centre affiliates and one for those not affiliated) were estimated within each balanced group. Of course, in other applications of our proposed method, the more common comparison-of-means between treatment and comparison cases may be the preferred approach, but in this case survival analysis is preferable. The survival analysis was carried out on the SAS system computing nonparametric estimates of the survivor function by the product-limit method (also known as the Kaplan-Meier method).

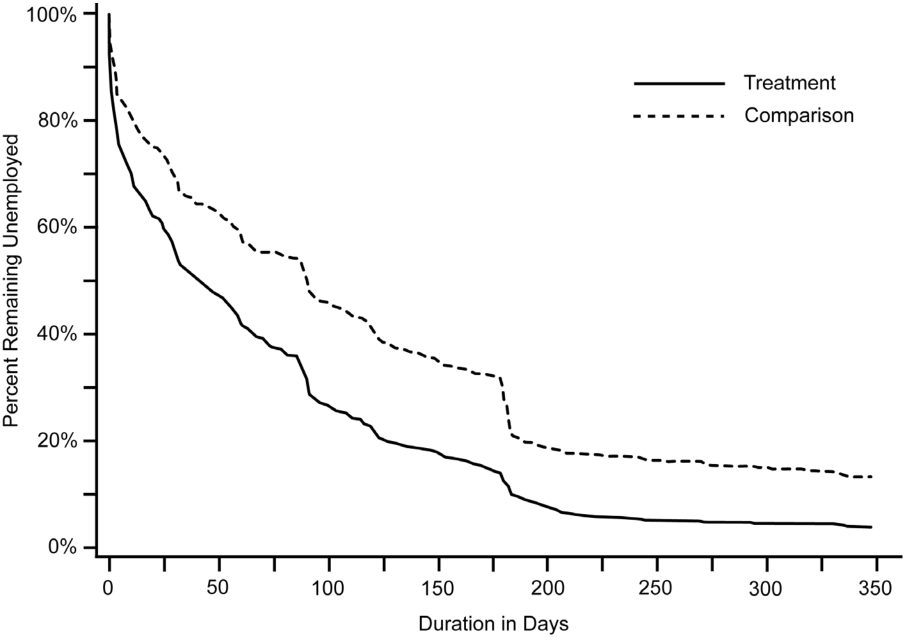
Table 3
Balance and Effects, by Cluster

Cluster	Treatment (%)	Comparison (%)	GI	Interval	Balance	Log-rank *p*-value	Wilcoxon *p*-value
1	0.0	100.0		no common support			
2	37.8	62.2	0.236	(0;0.261)	yes	0.17	0.04
3	39.3	60.7	0.013	(0;0.072)	yes	<0.01	<0.01
4	93.2	6.8	0.045	(0;0.118)	yes	0.24	0.89
5	36.2	63.8	0.043	(0;0.063)	yes	<0.01	<0.01
6	20.9	79.1	0.137	(0;0.225)	yes	0.84	<0.01
7	28.8	71.2	0.013	(0;0.058)	yes	0.79	0.92
8	30.7	69.3	0.010	(0;0.046)	yes	<0.01	<0.01
9	28.1	71.9	0.010	(0;0.102)	yes	0.23	0.26
10	30.1	69.9	0.002	(0;0.026)	yes	<0.01	<0.01
11	41.8	58.2	0.016	(0;0.106)	yes	<0.01	<0.01
12	26.5	73.5	0.017	(0;0.035)	yes	<0.01	<0.01
13	0.0	100.0		no common support			
14	47.7	52.3	0.058	(0;0.094)	yes	0.36	0.19
15	24.7	75.3	0.076	(0;0.104)	yes	<0.01	<0.01
16	21.3	78.7	0.025	(0;0.137)	yes	0.11	0.23
17	9.1	90.9	0.098	(0;0.125)	yes	0.95	0.14
18	44.6	55.4	0.069	(0;0.130)	yes	<0.01	<0.01
19	52.9	47.1	0.018	(0;0.049)	yes	<0.01	<0.01
20	42.0	58.0	0.214	(0;0.198)	no		
21	0.0	100.0		no common support			
22	100.0	0.0		no common support			
23	0.0	100.0		no common support			
24	0.0	100.0		no common support			
25	51.8	48.2	0.061	(0;0.116)	yes	<0.01	<0.01
26	75.4	24.6	0.067	(0;0.072)	yes	<0.01	<0.01
27	3.3	96.7	0.170	(0;0.078)	no		
28	18.9	81.1	0.049	(0;0.081)	yes	0.34	0.09
29	33.3	66.7	0.066	(0;0.084)	yes	<0.01	<0.01
30	24.2	75.8	0.067	(0;0.103)	yes	<0.01	<0.01

Note. We ran two tests of equality of the survivor functions between groups for each balanced cluster: the Log-rank and the Wilcoxon. For each test we report *p*-values, indicating whether the survival curves are statistically different between the cluster's treatment and comparison cases.

Consider two clusters in particular, numbers 19 and 28, as illustrative examples. First, for cluster 19, as shows in Figure 3, the Job Centre's program works at shortening the duration of unemployment. For this subgroup, both Wilcoxon and Log-rank tests (reported in the notes to Figure 3) let us conclude that the disadvantaged status of the "weaker" segments of the workforce is compensated by the intervention, as they find jobs more quickly than their comparison group counterparts, after accounting for bias.

Figure 3
Survival Curves of Job Start-ups, by Treatment Status, Cluster 19

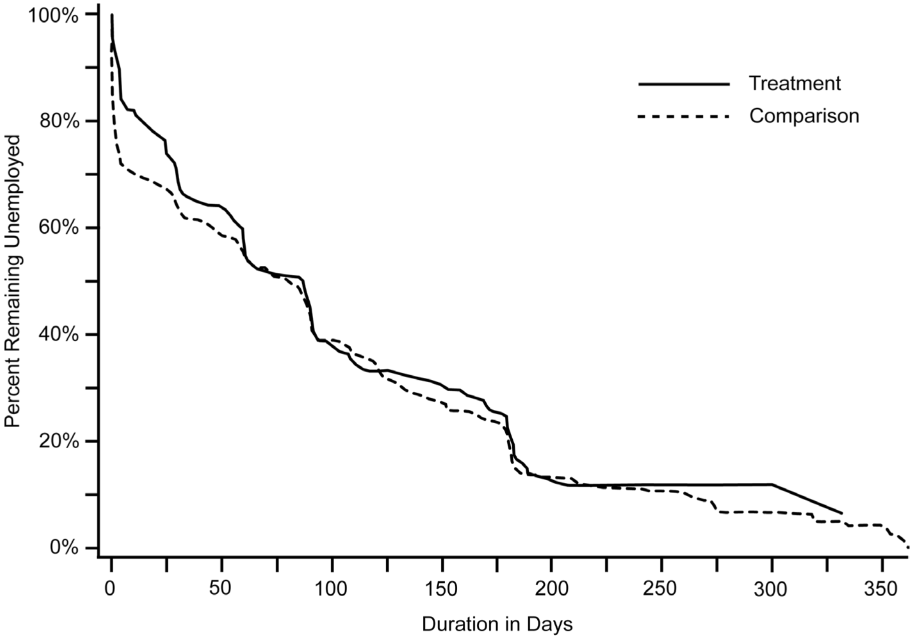


Note. We ran two tests of equality of the survivor functions between groups: the Log-rank (chi-square = 120.87; $p < 0.01$) and the Wilcoxon (chi-square = 100.35; $p < 0.01$). These both support the conclusion that the curves are statistically different from one another.

In contrast, for cluster 28, the Job Centre's programs are not effective at supporting disadvantaged workers' in overcoming labor market barriers (see Figure 4). Both the Wilcoxon and Log-rank show that being registered with the Job Centre makes no difference.

Table 3 shows that in 13 of the 22 balanced clusters there is a statistically significant difference in the duration of employment. Among these, 9 clusters (which comprise about 30% of the observations) demonstrate that the Bolognese Job Centres improved disadvantaged workers' employment outcomes. We note that it is not the intent of this article to explore the substantive meaning of the Bolognese Job Centres' effectiveness, though in most other practical applications it likely matters. In brief, our analysis of these data shows that the Bolognese Job Centres improved disadvantaged workers' employment outcomes in 30% of cases. Although these services are not sufficient to help all disadvantaged workers in overcoming their labor market barriers, we can say with confidence that the program works well some of the time.

Figure 4
Survival Curves of Job Start-ups, by Treatment Status, Cluster 28



Note. We ran two tests of equality of the survivor functions between groups: the Log-rank (chi-square = 0.91; *p* = 0.34) and the Wilcoxon (chi-square = 2.95; *p* = 0.09). The Log-rank test supports the conclusion that the curves are statistically equivalent, and the Wilcoxon test is on the edge of also supporting the conclusion that the curves are statistically equivalent.

DISCUSSION AND CONCLUSION

This article proposes an algorithmic approach to expunging selection bias from non-experimental data, thereby facilitating estimation of unbiased treatment effects. Our three-step process involves first identifying whether bias exists, then performing a cluster analysis on MCA coordinates (also known as the “tandem approach”), and finally comparing treatment and comparison cases within balanced clusters to estimate program impacts. We illustrate this analysis by using data from Bolognese Job Centres and learn that their activity is not sufficient to compensate the disadvantaged status of the “weaker” segments of the Italian workforce in all cases, but it is effective at accelerating their move out of unemployment for about one third of registrants. We also learn that it is possible to evaluate policies with a completely model-free procedure. This can enable policy makers to better understand—without the confluence of selection bias—the influences of their policies on various subpopulations.

We recognize the criticism that no amount of statistical adjustment can compensate for poor design, but the fact of the matter is that only rarely do we have ideal data, collected through an ideal evaluation design. Instead, we must make the most of information we have, recognizing limitations along the way. Our proposed analytic strategy is another in a toolbox that evaluators can use to assess the effectiveness of programs. We assert that a main strength of the approach proposed here is that it does not require subjective judgement about what variables to control for and instead relies on the existing variability with the data to determine the extent and direction of bias, using that information to restructure the data; in turn, comparing treated and untreated cases results in an estimate of impact that minimizes bias from selection.

Although our analytic process generates results that have internal validity, we recognize that—depending on the extent of purging required in Step 3—external validity may be limited. Specifically, we note in our application that we exclude 20% of the observations because they do not have a matched case (or cases) for the impact estimation. These mismatches mostly arise because of having no treatment or comparison cases in a resulting cluster, but they also arise in instances where selection bias has not been fully eliminated. As a result, our impact analysis considers just 80% of our observations, and this may limit in some way the generalizability of results to the population of interest. It is uncertain whether being confident about the causal inference among 80% of one’s sample is better than being insecure about the causal inference within one’s full sample. We would argue that the tradeoff of having a slightly smaller sample, and perhaps limiting generalizability, is worth achieving the greater confidence in impact estimates’ being unbiased by selection.

The practical advantage of our approach is that the heterogeneity of treatment effects, if present, is taken into account. As prior research has shown, computing an overall average treatment effect for a heterogeneous treatment group may obscure important impacts among subgroups (e.g., Peck, 2003), or overall impacts may be deemed insignificant when they actually are an accumulation of positive and negative effects among various subgroups (e.g., Bos, 1995). Similarly, as Heckman, Smith, and Clements (1997) discuss, value exists even when potential participants simply have the additional options that social programs confer, regardless of whether program impacts are favourable. Although this article does not aim to substantially inform the literature on the effects of job training programs, our findings do support the conclusion that the Bolognese model for assisting disadvantaged workers through Job Centres have varied impacts, with positive effects occurring about 30% of the time after eliminating the influence of selectivity within the population. Instead, this example aims primarily to demonstrate how to apply this new approach to estimating program impacts in instances where data are biased by selection. For the reason mentioned above, the procedure could be useful for a more general subgroup analysis since it helps discover for whom treatment works best.

It is not our intent here to provide proofs of this proposed approach's process of reducing selection bias; instead, we provide a brief introduction to the concepts and a simple illustration of our proposed approach. Our main goal has been to introduce a strategy for making causal inference from nonexperimental data when selection to treatment may otherwise bias impact estimates. Taking an algorithmic approach, we overcome persistent problems of researcher subjectivity in influencing analytic results. By synthesizing key concepts introduced in Peck (2005), D'Attoma (2009), and Camillo and D'Attoma (2010), we measure unbiased treatment effects where conditioning results are insignificant and apply an algorithmic clustering approach to devise unbiased treatment-control comparisons. We believe this approach holds promise as a powerful tool for evaluators.

We hope to whet the appetite of evaluators and researchers interested in the problem of selection bias and encourage them to consider this strategy in future work. This atheoretical approach has applications in many fields, with program evaluation being an important one. If researchers can strip their data of bias in the selection to treatment, then we can more confidently report unbiased effects of programs that matter for private and public good.

ACKNOWLEDGEMENTS

We acknowledge the useful input of participants in the ASU School of Public Affairs Research Colloquium, and we thank Cynthia Interdonato for research assistance.

NOTES

1. In applied mathematics, the "moment of inertia" is the integral of mass times the squared distance to the centroid (Greenacre, 1984).

2. For a comprehensive description of this method, computational details, and its applications, refer to Lebart, Morineau, and Warwick (1984), and for problems in the presence of a conditioning variable, refer to Camillo and D'Attoma (2010).

REFERENCES

Agodini, R., & Dynarski, M. (2001). Are experiments the only option? A look at dropout prevention programs. Unpublished manuscript.

Alboni, F., Camillo, F., & Tassinari, G. (2007). *Il dualismo del mercato del lavoro e la transizione da lavoro temporaneo a lavoro a tempo indeterminato in provincia di Bologna*. Atti della Commissione d'Indagine sul Lavoro. Consiglio Nazionale dell'Economia e del Lavoro. Camera dei deputati. Senato della Republica.

Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research. In R.P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 160–189). Oxford, UK: Blackwell.

Barnow, B.S. (1987). The impact of CETA programs on earnings: A review of the literature. *Journal of Human Resources, 22*(2), 157–193.

Bloom, H.S. (Ed.). (2005). *Learning more from social experiments: Evolving analytic approaches*. New York, NY: Russell Sage Foundation.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231.

Bos, J.M. (1995). The labor market value of remedial education: Evidence from time series data on an experimental program for school dropouts (Doctoral dissertation, New York University). Retrieved from *Dissertation Abstracts International* (56, 4139).

Camillo, F., & D'Attoma, I. (2010). A new data mining approach to estimate causal effects of policy interventions. *Expert Systems with Applications, 37*, 171–181.

Cox, D.R., & Oakes, D. (1984). *Analysis of survival data*. London, UK: Chapman and Hall.

Cox, D.R., & Wermuth, N. (1998). *Multivariate dependencies: Models, analysis and interpretation*. Boca Raton, FL: Chapman and Hall.

D'Attoma, I. (2009). A partial dependence factorial analysis to deal with selection bias in observational studies (Doctoral dissertation, University of Bologna). Retrieved from <http://amsdottorato.cib.unibo.it/1484/>

Dehejia, R.H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*(448), 1053–1062.

Dehejia, R.H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics, 84*(1), 151–161

Escofier, B. (1988). Analyse des correspondances multiples conditionelle. In E. Diday (Ed.), *Data Analysis and Informatics* (pp. 333–342). North Holland, Amsterdam: Elsevier Science.

Estadella, J.D., Aluja, T., & Thiò-Henestrosa, S. (2005). Distribution of the inter and intra inertia in conditional MCA. *Computational Statistics, 20*(3), 449–463.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. London, UK: Academic Press.

Heckman, J.J., Ichimura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies, 64*(4), 605–654.

Heckman, J.J., Smith, J., & Clements, N. (1997). Making the most out of program evaluations and social experiments: Accounting for heterogeneity in program impacts. *Review of Economic Studies, 64*(4), 487–535.

Kungl Vetenskapakademien, The Royal Swedish Academy of Sciences, The Sverige Riksbank Prize in Economic Sciences in Memory of Alfred Nobel. (2000). News release, October 11, 2000.

Lebart, L., Morineau, A., & Warwick, K. (1984). *Multivariate descriptive statistical analysis.* New York, NY: J. Wiley.

Luellen, J.K., Shadish, W.R., & Clark, M.H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29*(6), 530–558.

OECD. (2007). *OECD employment outlook*. <www.oecd.org/publishing/corrigenda>

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York, NY: Cambridge University Press.

Peck, L.R. (2003). Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice. *American Journal of Evaluation, 24*(2), 157–187.

Peck, L.R. (2005). Using cluster analysis in program evaluation. *Evaluation Review, 29*(2), 178–196.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 7*, 34–58.

Rubin, D.B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26*(1), 20–36.

Rubin, D.B., & Waterman, R.P. (2006). Estimating the causal effects of marketing interventions using propensity score methodology. *Statistical Science, 21*(2), 206–222.

Thoemmes, F.J. (2009). *The use of propensity scores with clustered data: A simulation study* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database (UMI No. 3380671).

Wilde, E.T., & Hollister, R. (2002). *How close is close enough? Testing non-experimental estimates of impact against experimental estimates of impact with education test scores as outcomes* (Discussion paper no. 1242-02). Madison, WI: Institute for Research on Poverty.

Appendix
Computing Global Imbalance and Its Significance Test

This technical appendix explains the computation of the measure of Global Imbalance (GI) and the imbalance test we propose using to ascertain whether selection bias (a) poses a problem for an evaluation's impact analysis and (b) has been eliminated from cluster subgroups.

The GI Measure

D'Attoma (2009) reports that the between-group inertia of a cloud of units denotes the GI measure expressed as:

$$GI = I_b = \frac{1}{Q}\sum_{t=1}^{T}\sum_{j=1}^{J_Q}\frac{b_{tj}^2}{k_{.t}k_{.j}} - 1$$

where

Q denotes the number of baseline covariates introduced in the analysis

T denotes the number of treatment levels;

J_Q denotes the set of all categories of the Q variables considered;

b_{tj} is the number of units with category $j \in J_Q$ in the treatment group $t \in T$;

$k_{.t}$ is the group size $t \in T$; and

$k_{.j}$ is the number of units with category $j \in J_Q$.

The GI measure is the result of using Conditional MCA (Escofier, 1988) that allows one to quantify the between-group inertia. Such a measure originates from the consideration that when the dependence between **X** and **T** is outside the control of researchers, displaying the relationship among them on a factorial space represents a first step for discovering the hidden relationship. In fact, if dependence between **X** and **T** exists, any descriptive factorial analysis may exhibit this link.

A conventional method dealing with the factorial decomposition of the variance related to the juxtaposition of the **X** matrix and the **T** variable is Multiple Correspondence Analysis (MCA) framework.[2] Given that the variability (inertia) of a data matrix can be decomposed into eigenvalues and eigenvectors, and referring to MCA for the study of the relationship between variables and of the structure induced by variables on the population, the presence of a conditioning variable (**T**) will strongly influence the structure of the matrix decomposition process. Hence, a conditional analysis could be useful in order to isolate the part of the variability of the **X**-space due to the assignment mechanism. Conditioning applied to problems arising from the dependence between categorical covariates and an external categorical variable was first studied by Escofier (1988) with the resulting Conditional Multiple Correspondence Analysis (MCA_cond).

Referring to Huygens' overall inertia decomposition of total inertia ($\mathbf{I}_T$) as within-groups ($\mathbf{I}_W$) and between-groups ($\mathbf{I}_B$), MCA_cond consists in a factorial decomposition of the within-group inertia. In turn, MCA_cond could be also considered as an *intra* analysis since the inertia induced by the conditioning variable (**T**) is not taken into account. Specifically, an *inter-group* analysis considers the relative position of groups, whereas an *intra-group* analysis detects and describes differences between units within each group by not considering the effect due to the partition's structure. In the evaluation context, this structure is induced by the nonrandom selection mechanism. An intra-analysis allows measuring the influence of conditioning, which means, as reported in Camillo and D'Attoma (2010), obtaining a measure of comparability between treatment groups.

This method especially works in the presence of categorical covariates. Eventually, continuous variables could be transformed into categorical by dividing them in classes. The need to work with categorical covariates stems from the consideration that, as reported in Cox and Wermuth (1998), in the social sciences, background knowledge tends to be qualitative.

The key result of using MCA_cond is represented by the quantified "Between-group Inertia" ($\mathbf{I}_b$). The *no omitted variable bias* assumption underlying the approach assumes a crucial role and thus must be emphasized. The assignment mechanism is assumed to be known, which means that the **X** matrix includes all baseline variables associated with both the treatment assignment and the observed outcome.

The Imbalance Test

To determine the significance of the detected imbalance, we perform an Imbalance test. We specify the null hypothesis of no dependence between **X** and **T** as

$$\mathbf{H}_0 : \mathbf{I}_W = \mathbf{I}_T$$

To establish an interval of plausible values for $\mathbf{I}_B$ under the null hypothesis, we use results obtained by Estadella, Aluja, and Thiò-Henestrosa (2005), who have studied the asymptotic distribution function of $\mathbf{I}_B$. Once the distribution of the between group inertia as

$$\mathbf{I}_B \approx \frac{\chi^2_{(T-1)(J-1)}}{\mathbf{nQ}}$$

has been derived, the authors have established the interval of plausible values for GI defined as

$$GI \in (0, \frac{\chi^2_{(T-1)(J-1),\alpha}}{nQ})$$

Specifically, if the GI calculated on the specific dataset is outside the interval, then the null hypothesis of no dependence between **X** and **T**[3] is rejected and data are deemed unbalanced. Simulation results show that where the test detects balance, unbiased estimates of the ATE are obtained (D'Attoma, 2009).

Dr. Laura R. Peck is an Associate Professor in the School of Public Affairs at Arizona State University. Her research focuses on how social welfare policies affect family well-being and on program evaluation methodology.

Dr. Ida D'Attoma is a Research Fellow in the Department of Statistical Sciences, University of Bologna (Italy). Her research interests include micro data mining, causal inference in observational studies, subgroup analysis, and new methods for public and private program evaluation.

Professor Furio Camillo is an Associate Professor of Business Statistics and Data Mining in the Department of Statistical Sciences, University of Bologna (Italy). He studies the applications of data mining in private and public organizations in the areas of marketing, customer relationship management, and policy evaluation.