The Leadership Quarterly 34 (2023) 101678

Contents lists available at ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

Full length article

Causal inference with observational data: A tutorial on propensity score analysis

Kaori Narita [a,*], J.D. Tena [a,b,c], Claudio Detotto [c,d]

[a] University of Liverpool Management School, Liverpool L69 7ZH, United Kingdom
[b] Università degli Studi di Sassari, Sassari 07100, Italy
[c] Centro Ricerche Economiche Nord Sud (CRENoS), Sassari 07100, Italy
[d] Università di Corsica Pasquale Paoli, LISA UMR CNRS 6240, Corte 20250, France

ARTICLE INFO

Keywords:
Causality
Propensity score
Leadership succession
Observational data
Football

ABSTRACT

When treatment cannot be manipulated, propensity score analysis provides a useful way to making causal claims under the assumption of no unobserved confounders. However, it is still rarely utilised in leadership and applied psychology research. The purpose of this paper is threefold. First, it explains and discusses the application and key assumptions of the method with a particular focus on propensity score weighting. This approach is readily implementable since a weighted regression is available in most statistical software. Moreover, the approach can offer a "double robust" protection against misspecification of either the propensity score or the outcome model by including confounding variables in both models. A second aim is to discuss how propensity score analysis (and propensity score weighting, specifically) has been conducted in recent management studies and examine future challenges. Finally, we present an advanced application of the approach to illustrate how it can be employed to estimate the causal impact of leadership succession on performance using data from Italian football. The case also exemplifies how to extend the standard single treatment analysis to estimate the separate impact of different managerial characteristic changes between the old and the new manager.

1. Introduction

Causal claims are present in most empirical research reported in the leadership literature. For example, analysts are interested in knowing the consequences of rewards (Fest, Kvaløy, Nieken, & Schöttner, 2021), traits (Rockey, Smith, & Flowe, 2021; Kiss, Cortes, & Herrmann, 2021), emotions (Sy, Horton, & Riggio, 2018) or previous experience (Zhang, Zhang, & Jia, 2021; Hopp & Pruschak, 2020). However, while randomisation provides a failsafe way to provide causal evidence, it is not always possible in social science. In particular, it could be challenging to operationalise complex constructs related to leadership in laboratory settings (Wofford, 1999) or, in some cases, to find situations in which key variables such as perceptions, choice, emotions or behaviours are quasi-randomised in natural experiments. Therefore, non-experimental designs are sometimes presented as the only feasible way to conduct research in social science. In this setting, propensity score analysis (PSA) allows for counterfactual comparisons under the strong ignorability assumption, which implies that conditional on observable variables, the potential outcomes are independent of treatment[1] status (Rosenbaum & Rubin, 1983). The application of PSA relies on the estimation of the probability of receiving treatment, or propensity score (PS). The two most common PSA approaches are propensity score matching (PSM) and propensity score weighting (PSW). They differ in the way they transform the sample to be used in causal analysis. While PSM uses PSs to form analogous treated and untreated observations, dropping non-matched observations, PSW uses all individuals in the original sample but weights them according to their PSs.

Despite some notable exceptions (see, for example, Vitanova (2021), Li et al. (2021)), PSA has been elusive in leadership and management research (Connelly, Sackett, & Waters, 2013; Schmidt & Pohler, 2018). This apparent absence of interest was highlighted in Li (2013, p. 209): "To my knowledge, no publications in the management field have implemented the PSM in an empirical setting, yet other social science fields have empirically applied the PSM", and Connelly et al. (2013, p. 416): "…most organizational researchers

[*] Corresponding author.
E-mail addresses: knarit@liverpool.ac.uk (K. Narita), jtena@liverpool.ac.uk (J.D. Tena), detotto_c@univ-corse.fr (C. Detotto).

[1] The word "treatment" originates from medical trials, where a certain treatment is given to the treated group, and no treatment (or a placebo) is given to the control or untreated group.

https://doi.org/10.1016/j.leaqua.2023.101678
Received 5 November 2020; Revised 19 January 2023; Accepted 8 February 2023
Available online 27 February 2023
1048-9843/© 2023 The Author(s). Published by Elsevier Inc.

who conduct quasi-experiments are generally not familiar with propensity scoring and have not generally considered using this technique in their research." Li (2013) and Connelly et al. (2013) provide comprehensive and insightful introductions to these methods for management scholars. However, almost one decade later, PSA is still rarely used in either the management or applied psychology literature. In this respect, Schmidt and Pohler (2018) indicate that econometrics, or statistical methods developed/used in economics, have been somewhat separated from other social sciences and underutilisation of PSA is perhaps one example of such. They also attribute the lack of popularity of PSA to the late arrival of statistical packages to deal with non-binary treatment variables, which have only recently become available. In this setting, an analyst must be aware of how counterfactual observations are defined for each treatment level. Thus, even with the help of statistical packages, an understanding of sophisticated automated algorithms and programming knowledge would still be required to interpret the estimates correctly.

This paper supplements the previous tutorials in three ways. First, we explain practical issues associated with the application of PSA in management whilst primarily focusing on the application of PSW. This method was initially proposed by Imbens (2000) and has been used in a variety of contexts, see Wooldridge (2010). PSW uses the inverse of the PS as a weight to apply to each treated unit and the inverse of one minus the PS as the weight to apply to each control unit (Imbens, 2000). Rather than relying on statistical packages with matching algorithms, the implementation of PSW only requires the application of weighted linear regression, which is readily available in most statistical software. Despite the simplicity that PSW can offer, most of the previous applications related to propensity scores are in matching (Thoemmes & Kim, 2011).

The second aim of the paper is to show and discuss research examples in the recent literature in management and applied psychology where PSA is used. All the examples postdate, and were therefore not included in, Li (2013) and Connelly et al. (2013). This review allows us to assess the use of PSA (and PSW, in particular) in the field, highlight its main assumptions, and also focus on rather advanced topics (e.g., PSW with non-binary treatment).

Finally, our third purpose is to provide an advanced practical example of how PSW can be used to study a leadership topic. In particular, we estimate the consequences of involuntary within-season managerial change in top-tier Italian football (*Serie A*) during seasons 2004/2005–2017/2018. Two aspects of this tutorial case are of particular relevance for leadership and management researchers. First, the example is written as a guide to implementing a double robust procedure that uses both PSW and regression adjustment to mitigate bias due to observables (Funk et al., 2011). In the standard PSW procedure, the treatment effect can be estimated within the weighted regression framework, where the weights are based on the estimated PSs, in order to control for the pre-treatment differences between clubs which dismissed managers and those which did not. In the weighted regression model where an outcome variable is regressed on a treatment variable, additional factors that can affect the outcome can also be included. However, an important limitation of PSW is that it is very sensitive to misspecification of the PS model (Freedman & Berk, 2008; Stone & Tang, 2013). Moreover, PSW does not perform well with small samples (Raad, Cornelius, Chan, Williamson, & Cro, 2020). Thus, to account for these concerns, our tutorial example employs a double robust procedure that increases protection against model misspecification by including the determinants of PSs in the weighted regression (Funk et al., 2011).

Another relevant feature of the tutorial is that it adapts the approach to deal with multidimensional treatment in PSW. In particular, we extend the analysis by considering leadership succession as simultaneous changes in the different dimensions of managerial characteristics (i.e., characteristics related to age, experience, association with the organisation, most recent employment status, and background). Our analysis shows that a positive outcome is expected following particular managerial characteristic changes. This highlights the importance of considering the different dimensions in which treatment is operationalised by management researchers.

The paper proceeds as follows. The following section explains the principles of PSA and how to conduct this type of research. Section 3 presents and discusses examples of the use of PSA in recent management and applied psychology research. Section 4 provides the illustrative case on the causes and consequences of head coach turnovers in Italian football. Finally, we offer ideas for future work and some concluding remarks.

2. Principles of propensity score analysis

2.1. The strong ignorability assumption

Causal inference would be straightforward in an ideal situation where we could observe the outcome of a subject i when receiving the treatment, $Y_i(1)$, *and* not receiving the treatment, $Y_i(0)$. The causal effect for unit i would be defined as:

$$c_i = Y_i(1) - Y_i(0). \tag{1}$$

Note that we could recover c_i if we were able to observe outcomes under each of the two scenarios (i.e. receiving and not receiving treatment) for an individual i. In reality, however, one individual is *either* treated *or* untreated. That is, the *observed* outcome for individual i can only be *either* $Y_i(1)$ (if i is treated) *or* $Y_i(0)$ (if i is untreated), hence c_i is unidentified (Holland, 1986).

Nevertheless, if treatment is randomly allocated, we can obtain an unbiased estimate of the average treatment effect (ATE) by comparing the average outcomes for treated and untreated groups in the trial sample as treatment is unrelated to each person's attributes and, therefore, independent of the potential outcomes $(Y(1), Y(0))$ (Fisher, 1935).

In observational studies, however, treatment allocation is unlikely to be random. For example, low-performing students are more likely to seek out test coaching than high-performing ones (Connelly et al., 2013). Similarly, Schmidt and Pohler (2018) note that observed employee satisfaction affects the level of interest in high-performance work systems investments. Since the very factors that sort individuals into treated and control groups can also influence an outcome of interest, we cannot attribute the differences in the outcome between the two groups to the pure effect of an intervention. That is, a direct comparison between treated and control groups is subject to selection bias. PSA is a tool to deal with such bias when treatment determinants can be observed.

To make causal inference possible, Rosenbaum and Rubin (1983) pointed out the need to assume strong ignorability, which requires the fulfilment of the following two conditions:

$$(Y(1), Y(0)) \perp T \mid X, \tag{2}$$

$$0 < \Pr[T = 1 \mid X] < 1 \tag{3}$$

Expression (2) is the unconfoundedness assumption which states that *potential* outcomes (Y(1), Y(0)) are not affected by (or are independent of) treatment assignment (*T*), conditional on a set of observable confounders (*X*), i.e. variables that influence treatment allocation and outcome. This property (also referred to as ignorability, conditional independence, or selection on observables) is fundamental to the statistical estimation of causal effects. For this condition to be fulfilled, it is necessary to assume that there are no unobservable variables simultaneously affecting the treatment assignment and the outcome variable. Therefore, the PSA relies on the strong assumption that there are no unmeasured confounding variables (Lee, Lessler, & Stuart, 2011; Shang & Rönkkö, 2022), and that all relevant confounders are included in the propensity score model (Emsley, Lunt, & Dunn, 2008).[2] However,

[2] Unobserved confounders are often referred to as "omitted variables" in leadership and management literature (Antonakis, Bendahan, Jacquart, & Lalive, 2010; Larcker & Rusticus, 2010).

it is not possible to test directly whether treatment assignment is “ignorable” (Guo & Fraser, 2014). The reason is that we do not know the distribution of $Y_i(0)$ for those who received the active treatment and that of $Y_i(1)$ for those receiving the control. Thus, researchers must identify the appropriate covariates based on theoretical and empirical grounds. Condition (3) is the overlap assumption. It means that every individual has a positive probability of being assigned to the treated and control group conditional on X. Overall, under the strong ignorability assumption, even if randomisation is not possible, it is credible to remove pretreatment differences between the treated and the control subjects in a sort of virtual randomisation (Rosenbaum & Rubin, 1983).

We provide simple simulation exercises in the supplementary material (Appendix A), which emphasise the importance of the strong ignorability assumption and the use of PSW as a way to obtain a “virtual randomisation” of treatment allocation under such an assumption. In particular, these exercises show that PSA can attenuate bias due to treatment selection when we can observe relevant confounders and employ them in the PS estimation. However, ATEs can still be biased if we fail to include all such confounders. Therefore, it is important to bear in mind that PSA is suitable when we can observe confounders. If there are unmeasured confounders, this requires a different approach, such as an instrumental variable method (see Hamilton & Nickerson (2003) and Semadeni, Withers, & Trevis Certo (2014) for the review of the method in management and leadership). The following section explains each step in conducting PSA.

2.2. Steps in the analysis

The PS is the *ex-ante* probability of a treatment assignment conditional on a collection of observed baseline variables (Rosenbaum & Rubin, 1983), which is estimated via prediction models for treatment allocation. PSs can be used to identify individuals who are similar in terms of pre-treatment conditions but only differ in treatment assignment (treated or control). Based on this, different types of PSA can be applied to adjust a sample so that the covariates are more similar (“balanced”) between the treated and control groups, as though the treatment had been randomly allocated.

PSA typically comprises four steps, as illustrated in Fig. 1. In the first step, a PS model is specified as a function of observed variables related to pre-treatment conditions. Probit and logit models are the usual approaches to estimate treatment probabilities (Caliendo & Kopeinig, 2008). However, more advanced classification methods based on machine learning are also available (Lee, Lessler, & Stuart, 2010). Again, it is important to include a set of relevant covariates selected on theoretical and empirical grounds. It is generally advised that confounders that affect both treatment assignment and outcome should be included in the treatment assignment model, whilst variables that only affect treatment assignment but not directly the outcome should be left out (Austin, 2011; Heinze & Jüni, 2011).

The second step differs between PSM and PSW, the two different “virtual randomisation” strategies. The former employs an algorithm to find pairs of individuals in the treatment and control groups with similar PSs. Several alternative algorithms can be used for this purpose. As indicated in Fig. 1, PSM links n individuals in the treatment group to their closest m individuals in the control group according to their estimated PS. One of the most popular, nearest neighbour matching, finds one or more units with the closest PS within the control group for each treated individual (i.e. 1:1 or 1:m). The process is repeated until no observations are left in the treatment or control group. Other matching approaches, such as optimal matching, also exist. Optimal matching aims to minimise the average absolute distance across all matched pairs (Gu & Rosenbaum, 1993).

Matching algorithms become especially cumbersome in the case of multiple or continuous treatments. In the former case, it is still possible to estimate PSs using multinomial logit or probit models and make paired comparisons with a reference treatment group. For example, Hopp and Pruschak (2020) deal with this problem by separately estimating the effect of each treatment using PSM, and they explain that results are robust to a multinomial treatment estimation of PSs. A possible problem with this approach is that, because some observations are dropped during matching, each paired comparison may be based on different individuals. In the case of continuous treatment, Hirano and Imbens (2004) present a matching approach based on estimating the treatment dose rather than its PS.

Another potential issue with PSM is that it requires many individuals, especially in the control group. Moreover, certain matching schemes may not use a large number of observations (Stuart, 2010). Contrarily, PSW, in principle, retains all the observations (Guo & Fraser, 2014). A second advantage is the simplicity of PSW. In the PSW approach, a weight allocated to each individual is defined by the inverse of the estimated PS for a realised treatment status. Intuitively, a treated unit with a low probability of being treated is given a high weight, and a control unit with a high probability of being treated is also given a high weight. In doing so, the distribution of the *ex-ante* probabilities of being treated become similar across the treated and control groups, as though the treatment were assigned randomly. Therefore, the second step in PSW only involves obtaining these weights to be employed in a weighted regression, similar to the application of sample or survey weights commonly used in social sciences. Furthermore, this approach can be relatively easily generalised to multi-treatment cases, as in Schmidt and Pohler (2018) and Love, Lim, and Bednar (2017). While weighted regressions are readily implementable with most of the statistical software available, applying matching algorithms typically requires becoming familiar with specialist tools such as matchIt in R or psmatch2 in Stata.

The third step is common to PSM and PSW and consists of testing for balance in covariate distributions between the treatment and the control groups. The general idea of these checks is to compare differences between the treated and the control group before and after matching or weighting. Two common approaches are (1) the standardised bias, which assesses the distance in marginal distribution of the X variables, and (2) a two-sample t-test to check whether there are significant differences in covariate means for both groups (Rosenbaum & Rubin, 1985). If these tests are not completely successful, some remedial measures are advised, such as including interaction terms in the PS estimation (Caliendo & Kopeinig, 2008).

The final step in PSA consists of estimating the impact of treatment on the variable of interest. The treatment effects can be obtained by comparing the average outcomes between treated and control units within matched samples (PSM) or comparing the weighted average of outcomes between treated and control groups (PSW). Alternatively, one can estimate the treatment effect through multiple regression, again using matched samples with PSM and through a weighted multiple regression with PSW.[3] This regression has two purposes. First, it can be used as a double robust procedure where treatment determinants are included to further protect against the bias due to observables (Funk et al., 2011). Furthermore, it also allows controlling for additional factors that can potentially impact the response variable after treatment.

In general, two main definitions of treatment effects are considered: the average treatment effect (ATE) and the average treatment effect on the treated (ATT). The decision on which of the two causal effects are estimated depends on the researcher’s interest and the PSA method employed. For instance, consider a PSM design such that, for all treated individuals, the closest individual in the control group is matched. By averaging the differences in the outcomes of these two groups, we would estimate the ATT. However, by evaluating the

[3] As noted above, with PSW, weighted regression directly applies the weight defined in step 2. However, the weights obtained from PS estimates should not be confused with weights in survey sampling. While the former tries to address endogeneity in the treatment assignment, the latter is intended to adjust the sample data to reflect population attributes.

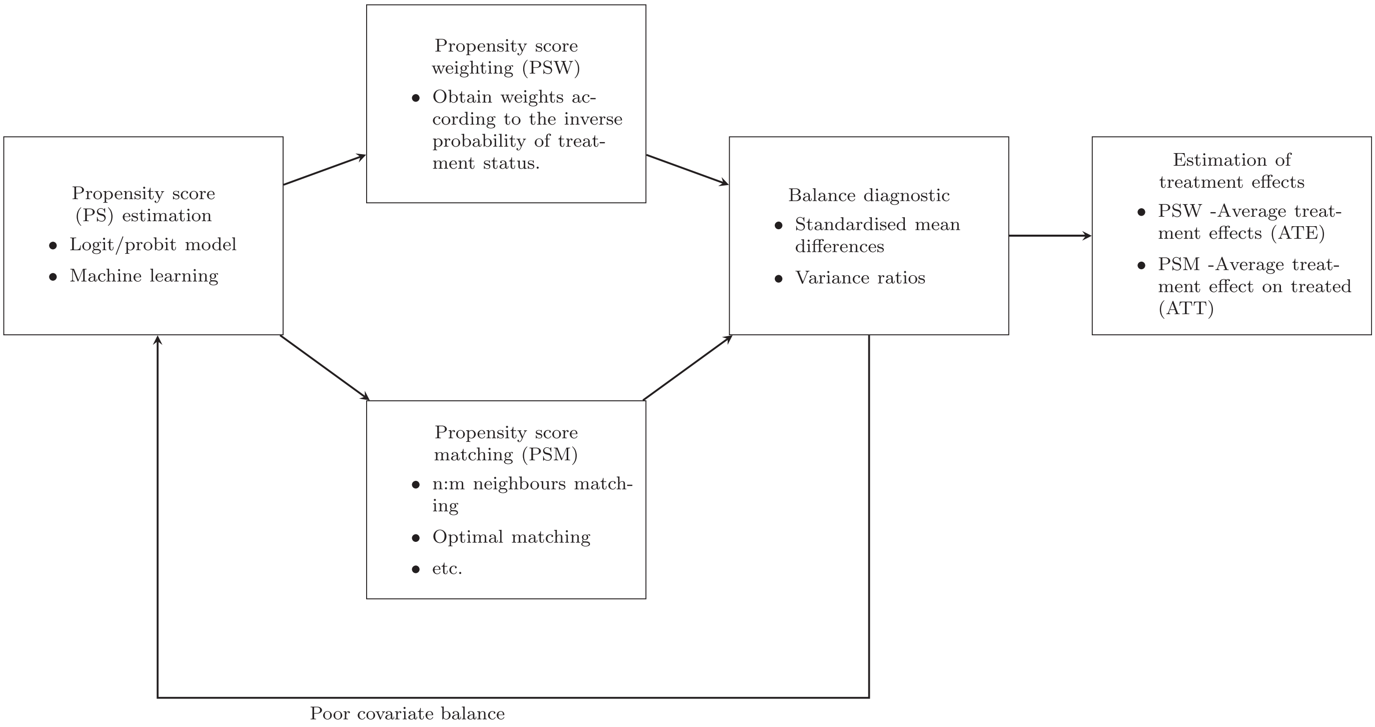
Propensity score (PS) estimation
• Logit/probit model
• Machine learning
Propensity score weighting (PSW)
• Obtain weights according to the inverse probability of treatment status.
Propensity score matching (PSM)
• n:m neighbours matching
• Optimal matching
• etc.
Balance diagnostic
• Standardised mean differences
• Variance ratios
Estimation of treatment effects
• PSW -Average treatment effects (ATE)
• PSM -Average treatment effect on treated (ATT)
Poor covariate balance

Fig. 1. Steps in propensity score analysis.

impact of treatment on the whole weighted sample, PSW provides an estimate of ATE.

2.3. Other validity concerns

The previous sections describe how PSA can attenuate selection bias due to observables and, under the assumption that there are no unmeasured confounders, estimate causal effects. However, other validity concerns may remain in the analysis even if this assumption is satisfied. A taxonomy of the most prominent causal threats can be found, for example, in Cook and Campbell (1976) and Crano, Brewer, and Lac (2014). Podsakoff and Podsakoff (2019) summarise the main validity threats and provide illustrations from the leadership literature. These concerns can be split into internal and external validity threats. Internal validity requires correctly attributing differences in the dependent variable to treatment variations. Podsakoff and Podsakoff (2019) identify potential validity threats due to selection, history, maturation, testing, instrumentation, regression, mortality and selection by maturation interactions.

Depending on the characteristics of the observational sample, some internal validity threats can be particularly relevant in PSA. In particular, the history threat is a consequence of external events affecting individuals over time between the impact of the treatment and the instant when the dependent variable is observed. Unlike history, maturation is not related to external events but to the way individuals evolve over time. For example, they may become older, more tired or less motivated than at the time of treatment. History and maturation threats increase the larger the length of time between the treatment and the measurement of the response variable(s) (Podsakoff & Podsakoff, 2019). For example, Hopp and Pruschak (2020) test the long-term consequences of educational decisions on earnings 11 and 50 years later. Also, Zhang et al. (2021) look at the biographical information of experienced executives to estimate the impact of having previous military experience on the frequency that his/her company engages in pollution-causing activities. History and maturation threats are relevant concerns in these circumstances. However, looking at short-term reactions of the dependent variable(s) could also be problematic if the analysis involves situations where the treatment requires some time before having its effect. For instance, some time is needed to assess the final impact of training on workers' productivity. Therefore, coping with this problem requires estimating the sensitivity of estimates to using different periods and controlling for all the possible factors affecting the dependent variable. These issues are particularly worrying when there is an interaction of selection with history and maturation.

However, even where these internal validity conditions are fulfilled, a fundamental question is the generalisability of causal effects in other settings. In particular, two main external validity concerns are (1) generalisability of operationalisations and (2) generalisability of results to other places and participant populations (Cook & Campbell, 1976; Crano et al., 2014). The validity of operationalisations concerns the correct identification of the treatment and response variables and the underlying relationship between them. A "treatment" could have many different meanings. In PSA, this concern also requires estimating the different impacts of different treatment intensities or subgroups in observational samples. For example, Schmidt and Pohler (2018), Love et al. (2017), and the tutorial case in Section 4 show how to conduct such analysis with PSW (see also Boivie, Graffin, Oliver, & Withers (2016) and Hopp & Pruschak (2020) for how to conduct such analysis using PSM).

3. PSA in management, psychology, and leadership research

Propensity scoring is still rarely used in the management, psychology, and leadership literature. To illustrate this issue, we explored the same top-tier journals surveyed by Antonakis et al. (2010) in their review of causal analysis: *Academy of Management Journal, Journal of Applied Psychology, Journal of Management, Journal of Organizational Behavior, The Leadership Quarterly, Organizational Behavior* & *Human Decision Processes*, and *Personnel Psychology*. We added *Strategic Management Journal* to this search as it is an FT50 journal that contains some examples of PSA in management.

Initially, for the purpose of comparison, we considered 4,330 abstracts in these journals from 2015 (two years after the publication of the two PSA tutorials by Li (2013) and Connelly et al. (2013)) to 2022[4]. In this search, the term "propensity score" appeared in 8 abstracts. Additionally, we accounted for the possibility that papers may have employed PS in causal analysis without necessarily using the term "propensity score" in the abstracts. Consistent with this possibility, we found 40 instances where the word "matching" appeared without "propensity score". However, in these cases, only three papers conducted PSA. This amounts to a total of 11 papers (0.25 per cent) that refer to PSA in the abstract, compared to, for example, 47 and 70 for "laboratory experiments" and "field experiments" respectively. In this group of 11 papers, we could only find three examples of PSW studies. However, only two of them use PSW in their core analysis because Rocha and Van Praag (2020) employ PSW as one of three alternative methods to deal with endogeneity in a robustness exercise.

To identify and discuss more specific examples of PSW in the extant management literature, in addition to the previous search, we account for the possibility that papers could still employ PSA without referring to it in the abstract. Thus, first, we searched the term "propensity score" in the text of the 4,330 articles.[5] Then, in a second step, we visually inspected the selected cases to identify 25 additional studies that conduct PSA as part of the main econometric analysis. It is worth noting that out of these 25 articles, only 4 of them conducted PSW (with the rest conducting PSM), suggesting that PSW is even more underutilised. In the remaining of the section, we discuss the six examples of PSW identified through abstract and main text search as above. Whilst some of these examples are not related to leadership (in particular, Examples 3, 4 and 6), they present interesting PSW applications that are worth reviewing.

In addition, in Appendix B of the supplementary material, we provide a brief review of eight examples of PSM (Chen, 2015; Boivie et al., 2016; Bechtoldt, Bannier, & Rock, 2019; Gupta, Mortal, Chakrabarty, Guo, & Turban, 2020; Li et al., 2021; Vitanova, 2021; Hopp & Pruschak, 2020; Ong, 2021) and one example that used PSW as an alternative method (Rocha & Van Praag, 2020) identified through the abstract search. Although the literature search strategy was precise and meticulous, it is still possible that some articles using PSW have been missed out in the literature search. This may be due to several factors, such as system errors on the Google Scholar portal not allowing some papers to be tracked, missing papers due to a delay in an online publication, and the use of different terms referring the PSW technique.

Example 1: The importance of CEO-CFO social interaction to explain outcomes for the CFO and the organisations

Background: Shi, Zhang, and Hoskisson (2019) examine the role of interactions involving chief executive officers (CEOs) and chief financial officers (CFOs) to explain outcomes for the CFO and organisations. More specifically, they measure the level of CEO-CFO verbal mimicry from common function words (e.g., articles, pronouns, auxiliary verbs, and conjunctions) observed in conference calls in the context of firm mergers and acquisitions. They denote this measure as CEO-CFO language style matching (CEO-CFO LSM). Using different regression analyses, they find that CEO-CFO LSM explains CFO compensation, the likelihood of the CFO becoming a board member, and the number and value of mergers and acquisitions.

Methodological design: To deal with the fact that the level of CEO-CFO LSM is not randomly assigned but selected by the CEOs, the authors implement a PSW analysis. First, they estimate a probit model predicting the probability of a firm having a high or low level of CEO-CFO LSM. They code it as a binary variable using the median value of CEO-CFO LSM. In the probit model, they include firm-level variables and previous information about the CFO. Then, they use the inverse of the PS calculated from the probit regression as a weight in regressions for different outcomes. The focus variable in these regressions is the level of CEO-CFO LSM, but the authors also control for other firm and CFO characteristics as well as other CEO-CFO similarities.

Strengths and limitations: The paper addresses a relevant question in the management literature, the role of social interaction in explaining firm outcomes. It presents the PSW regression to complement previous regressions that do not explicitly deal with the endogeneity of CEO-CFO LSM. One limitation is that the paper does not provide information about whether the application of PSW makes the sample more balanced in terms of observable variables. This is an essential consideration when interpreting PSW results. Another limitation is that transforming their continuous treatment variable (CEO-CFO LSM) into a binary one is arbitrary. Results could be different if another transformation rule were used.

Example 2: The role of leader behaviour in understanding the effect of HPWS on employee and consumer satisfaction

Background: Schmidt and Pohler (2018) use PSW to estimate the causal impact of high-performance work systems (HPWS) on employee and customer satisfaction using longitudinal survey data from a financial service organisation in Canada. They test the hypothesis that leadership behaviour confounds the relationships between HPWS and employee/customer satisfaction, where leader behaviour is measured by subordinates' perceptions of the leadership.

Methodological design: Given that the treatment variable is not binary but continuous, i.e. the level of HPSW, the paper employs a non-conventional PSW method. In particular, as Schmidt and Pohler (2018) indicate, transforming a continuous variable into a dichotomous variable consisting of treatment and control conditions is problematic. It requires arbitrary judgment that results in a loss of information and could generate model specification problems. Therefore, they use the covariate balancing propensity score for a continuous treatment proposed by Fong, Hazlettand, and Imai (2018). This procedure assigns a weight to each observation, minimising the association between treatment and covariates. Using these weights, they specify regression models to estimate the two-way causality between HPWS and employee and customer satisfaction and the role of leader behaviour as an omitted variable in the causal relationship.

Strengths and limitations: Overall, the paper provides an interesting example of the use of PSA to make individuals subject to different levels of treatment comparable in terms of observable variables. More importantly, it also provides a way to deal with a continuous treatment variable in causal analysis. A potential caveat of this analysis is that treatment allocation may depend on unobservable variables. Although the authors indicate that an instrumental variable analysis would be a way to tackle this concern, they do not pursue this approach as "it is very difficult to find a justifiable instrumental variable in survey-based research" (Schmidt & Pohler, 2018, p. 1013).

Example 3: How internet activism affects the speed of donations in firms

Background: Using information from 613 large publicly listed Chinese firms, Luo, Zhang, and Marquis (2016) study how internet acti-

[4] The search took place on 22/03/2022. We explored all publications from Scopus between 2015 and 2022 after removing editorials (93), errata (67) and one retraction. The terms "propensity score" and "matching" were employed to identify potential PS studies.

[5] Our search took place on Google Scholar on 09/04/2022. We explored the presence of the term "propensity score" within the document (abstract, main text and references) for all publications between 2015 and 2022 in the selected journals. By entering these search criteria in the Google Scholar database, a total of 79 papers were recorded and downloaded: 26 from the Academy of Management Journal, 6 from the Journal of Applied Psychology, 21 from the Journal of Management, 1 from the Journal of Organizational Behavior, 23 from The Leadership Quarterly, and 2 from Organizational Behavior and Human Decision Processes. We downloaded these papers manually for further inspection.

vism, and its interaction with other firm indicators, affected the speed of donations after the 2008 earthquake in the Sichuan Province of China.

Methodological design: The study uses continuous-time event history design to estimate how quickly companies reacted to the 2008 earthquake with donations. The dependent variable is the hazard rate of donation. Luo et al. (2016) employ a wide set of independent variables that include measures of internet activism, media coverage, reputation, the political status of top executives and indicators for state-controlled or belonging to a culpable industry[6]. Given that firms that donate and do not donate are not comparable, Luo et al. (2016) estimate the PS for donation prior to the earthquake using a probit model. In the second step, they adjust the event history regression through PSW. The paper does not provide detailed information about the PS specification or balance tests. However, a relevant aspect of the research is that they use weighted regression to estimate the impact of different independent variables on the hazard rate of donation.

Strengths and limitations: Luo et al. (2016) show the contribution of different sets of variables to the regression of the speed of donation by adding these variables in sequential steps. They also show that results are robust to employing an OLS regression and a Heckman regression model that corrects for potential selection bias in non-donating firms. As the authors acknowledge, the limitations of the study are related to the fact that relevant features of online media (such as the number of times an article was forwarded) or a wide range of online tactics are not included in the analysis. They are potentially confounding variables, and not including them may have resulted in an underestimation of the impact of Internet activism. Also, the paper does not report or mention balance tests. Nevertheless, Luo et al. (2016) provide a novel and interesting example of the use of PSW to study the determinants of firm donation decisions.

Example 4: The role of partners' administrative controls to explain knowledge transfer

Background: Devarakonda and Reuer (2018) analyse how partners' administrative controls in nonequity collaborations affect knowledge transfer across partners. They postulate that technology overlap and the value of the partners' knowledge drive the degree to which partners build upon each other's knowledge. They also hypothesise that this effect is moderated by steering committees.

Methodological design: Devarakonda and Reuer (2018) estimate the impact of a set of independent variables, including technology overlap and a steering committee indicator on cross-citations in publications by the client and R&D firms. For this analysis, they use pooled cross-sectional data from alliances in the biopharmaceutical industry. They address the problem that the choice of using a steering committee is not random by implementing a PSW analysis. Thus, they first estimate the PS for steering committees. Then, they weight observations with the inverse of the PS and estimate the determinants of cross-citations by the client and R&D firms in a negative binomial framework. It is worth noting that they follow a double robust approach as the outcome regression includes the confounders employed in the PS specification. Additionally, the outcome regression includes other covariates regarding experience and alliance citation potentially affecting the response variable.

Strengths and limitations: Devarakonda and Reuer (2018) employ a number of robustness exercises to study the effect of steering committees on knowledge flows in nontechnological areas, finding similar results for the R&D firm but not for the client firm. The paper also shows an interesting example of applying PSW to a case where the output regression is non-linear. They explain that their model does not control for the dynamic effects of the steering committee that could be just responding to an incipient problem of misappropriation of knowledge.

Example 5: The influence of CEOs on corporate reputation

Background: Love et al. (2017) study how CEOs influence corporate reputation. In particular, they hypothesise that companies whose CEOs receive more media attention will have a stronger reputation, and this effect will be stronger the more positive the amount of media attention is. They state that a stronger reputation can also be explained by CEOs having outsider standing or having received industry awards.

Methodological design: The authors test the hypotheses using separate models for each of the independent variables. They had two main issues to address. The first one is the potential endogeneity of the independent variable that motivates the use of PSA. The second problem stems from the fact that some of the independent variables are not dichotomous. The authors dealt with these two issues by using a weighting scheme based on the generalised propensity score technique (Imbens, 2000). Thus, in the first step, they run multinomial logit regressions to estimate PS for each category of the independent variable conditional to firm and CEO characteristics. They use specific control variables in each PS regression. Then, PSs are used to weight each category and estimate the impact of the different treatments on the measure of firm reputation (the dependent variable).

Strengths and limitations: The methodological part of the paper shows how to use PSW to conduct causal analysis in settings with non-dichotomous treatment variables. The article also shows that their results are robust to the use of time-series output regressions with fixed effects. As is common in PSA, a general concern is the endogeneity of the treatment variable because it might be explained by omitted variables. Love et al. (2017) address this issue using a cumulative count of awards as an instrumental variable. However, the authors acknowledge not being able to find valid instruments for media coverage and outsider status. Other concerns, also mentioned by the authors, are that the study uses a short time period (from 1991 to 1997) and that some relevant CEO characteristics could be omitted.

Example 6: How do political and executive ties affect the sell-off strategy of firms?

Background: Zheng, Singh, and Chung (2017) appraise the relevance of political and executive ties to affecting the sell-off strategy of firms in emerging markets. They also study how this effect is moderated by the capital market and how developed the legal system is.

Methodological design: They use a categorical indicator with three outcomes (sell-off, dissolution, or survival) as the dependent variable and indicators of political ties and institutional development as explanatory variables. Because firms with and without political ties are not comparable, the authors estimate the propensity to establish political ties by means of a probit regression. This PS regression includes five additional control variables not employed in the output regression that "may influence the formation of political ties but are not directly associated with sell-offs." (Zheng et al., 2017, p. 2021). Then, the second step uses the estimated PS to reweight the sample and estimate the likelihood of sell-offs employing a multinomial logit regression.

Strengths and limitations: The paper provides an interesting example of using PSW to conduct causal analysis in a multinomial logit output regression. In addition, the authors explore alternative explanations for the results. In particular, they test whether political ties lead to poorer firm performance, finding non-significant results. They further estimated the interaction between political ties and state ownership and find that state ownership decreases the effectiveness of political ties but not legislative ties. This shows how a treatment effect can be moderated by external factors. In the robustness exercise, the

[6] Firms belonging to a culpable industry are those subject to criticism for a number of reasons (such as suspected corruption, substandard construction or allowing executives to accumulate abnormally large fortunes).

number of political ties (as opposed to a dichotomous variable indicating at least one political tie) was included in the analysis, where their results remain similar. However, the paper does not provide details on how such analysis is conducted using PSW. One of the limitations acknowledged by the authors is that the study does not account for unofficial ties, such as family and social relationships. The inclusion of such ties could potentially strengthen informal ties. Therefore this could affect the estimated effects of political ties overall.

General discussion

The discussion above and the review of additional studies in Appendix B show how PSA has been used in management, psychology, and leadership outlets. Still, two main concerns can be mentioned. First, papers must provide enough detail about how the research is conducted. For instance, showing that weighting significantly improves covariate balance is essential in order to know whether PSA makes the treatment and control groups comparable in terms of observable variables. Some articles, however, failed to report such important details. A second concern is that, even if PSA is rigorously conducted, it might not be enough to infer causality. More specifically, some internal validity threats are also present in many of the examples due to unobserved pre-treatment differences, maturation or history and attrition, among others. The papers show different ways to deal with these concerns. One possibility is to check how changes in the methodological design within a given study affect results. Again, an important aspect of PSA is that it relies on the assumption that there are no unobserved confounders. In other words, it does not control for bias due to omitted variables. In cases where such variables are present, employing instrumental variables (Gupta, Han, Mortal, Silveri, & Turban, 2017; Hopp & Pruschak, 2020) is a way to control for the impact of omitted variables. However, the validity of instrumental variables lies in the credibility of the exclusion restriction (i.e. the instrument only affects the outcome variable through its effect on the endogenous covariate). Suitable instruments that satisfy this restriction are often lacking in the data set.

Another relevant approach is to estimate causal effects in a regression model that permits double control for treatment predictors and/or other determinants of the response variable (Vitanova, 2021; Boivie et al., 2016; Schmidt & Pohler, 2018; Love et al., 2017). This approach does not solve endogeneity problems associated with unobservable confounders but lessens misspecification concerns.[7]

A challenge in future research would be to adapt PSA to explore better how the treatment is operationalised in a multi-treatment setting. One possibility is to estimate the different impacts of different treatment levels rather than dichotomising the treatment variable. Another option is to decompose the treatment variable into different sub-treatments to study each effect. In this regard, the example in Hopp and Pruschak (2020) illustrate how one can implement such analysis using PSM. However, due to its simplicity, PSW provides an alternative way to deal with the multi-treatment extension as it only requires weighting observations according to the inverse of the PSs for each treatment level. Love et al. (2017), Schmidt and Pohler (2018), and the tutorial in the following section are examples of such an approach for non-binary treatments.

4. A tutorial on PSW: Leadership succession effects

In this section, as an illustrative and advanced example of PSW, we present some results based on leadership succession data. Leadership replacement is one of the crucial decisions that can shape the performance of an organisation. Given its relevance, the matter has attracted the attention of researchers from different fields and with diverse backgrounds and interests. For instance, Berns and Klarner (2017) provide a complete review of the factors affecting the impact of CEO succession in publicly traded firms, Farah, Elias, De Clercy, and Rowe (2020) extend their discussion to leadership changes in privately owned businesses and political organisations. Among these studies, the field of professional sports is particularly well suited to study leadership succession by offering stronger internal validity (Giambatista, Rowe, & Riaz, 2005; Rowe, Cannella, Rankin, & Gorman, 2005). Regarding this aim, event studies, the most common methodological approach in this context, requires a precise definition of event dates, confounding factors, and event windows (de Jong & Naumovska, 2016). Great interest among the public in professional sports means that the dates of and reasons for head coach replacements are widely covered by the media. Head coaches are interesting leaders to study since they occupy a role akin to that of a chief operating officer in a conventional firm in conventional firms (Hughes, Hughes, Mellahi, & Guermat, 2010).

Second, the firm’s objective, sporting success, is clearly defined, and such performance is frequently and regularly documented. Finally, we can clearly identify some confounding variables, such as the characteristics of a club and the difficulty of a match. In fact, using sports data is a recent trend in management studies (Fonti, Ross, & Aversa, 2022).

The following empirical example shows how to estimate the consequences of involuntary within-season managerial change in top-tier Italian football (*Serie A*) during seasons 2004/2005–2017/2018. An essential identification issue in such analysis is that managerial dismissal is not a random event. For example, it tends to occur particularly when a club is performing poorly. Therefore, estimation results can be biased if this issue is not properly accounted for. Numerous studies analyse this issue using regression models that include previous performance information among the regressors (Audas, Dobson, & Goddard, 2002; Tena & Forrest, 2007). However, a possible problem with regression analysis is that it is not informative on whether treated and control individuals are comparable in terms of observables. More recent research has employed matching methods to find comparable counterfactuals in terms of observable variables. For example, Muehlheusser, Schneemann, and Sliwka (2016) use previous performance as the matching variable. van Ours and van Tuijl (2016) consider control groups formed with counterfactual observations that followed a similar path of cumulative surprise[8] but where the clubs did not replace their manager.

Our tutorial example shows how to address the question of the effect of head coach replacement by employing PSW. The PS is estimated as a function of multiple variables related to indicators of recent match outcomes, relative performance compared to expectations, position in the league, and recent performance in other competitions. The method used in the exercise is a double robust estimator as we control for determinants of managerial dismissals in two regressions, one for treatment assignment, i.e., head coach replacement, and another for the outcome variable (Funk et al., 2011). Such an approach offers protection against misspecification as only one of the two specifications needs to be correct.

The second aim of the example is to show how to address a critical challenge faced by empirical researchers, the simultaneous estimation of the impact of multiple treatments. As discussed in the previous section, a limited number of papers consider a non-binary treatment (Hopp & Pruschak, 2020; Schmidt & Pohler, 2018). In this example, rather than just focusing on the aggregate impact of a head coach dismissal on future performance, we explain in addition how to estimate

[7] For example, Gupta et al. (2017), Hopp and Pruschak (2020), Vitanova (2021), Boivie et al. (2016) use PSM. A brief review of these studies is available in Appendix B of the supplementary material.

[8] In particular, the authors employ the difference between the actual number of league points won and the expected number of points according to the match outcome probabilities captured in betting odds, accumulated since the beginning of a season.

the effect of a set of changes in managerial characteristics. Again, the proposed setting is particularly appropriate for this type of analysis as the natural time for changing leadership in sports clubs is at the end of the season (Tena & Forrest, 2007). This implies that the possibility of selecting a new head coach, let alone each of their different characteristics, in a within-season football turnover is limited in terms of available candidates and time to reach an agreement. Nonetheless, we also explain in an advanced application in Section 4.7 how to adapt the PSW analysis to deal with the possible endogeneity of the similarity in characteristics between dismissed and new coaches.

Our proposed estimation is relatively simple, at least from a programming/software viewpoint, because it is built upon the standard regression analysis. This means that no statistical package specialised in PSA is required for this estimation. The interested reader can find the dataset and R codes used in this analysis in the research data available online.

4.1. Data

We collected club-match level data from the top tier of the Italian professional football league (*Serie A*) for seasons 2004/2005–2017/2018, which gives a total of 10,640 observations (5,320 matches). Throughout the season, each club competes against all others, once at their home stadium and once away. In each match, a club is awarded 3, 1, or 0 points for a win, draw, or loss, respectively. At the end of the season, the club with the highest accumulated points wins the championship title, whilst the three lowest-placed clubs are relegated to the lower-tier league (*Serie B*). The league publishes official match reports. They contain, for example, the names of each club in the match, the respective managers and the outcome of the match. Additional sources used are provided below, together with the descriptions of (1) the treatment variable, (2) variables that explain treatment assignment, and (3) outcome variables and additional control variables associated with the outcome.

4.1.1. Treatment variable

Our treatment variable $New\ coach_t$ takes the value 1 if $Head\ coach_t \neq Head\ coach_{t-1}$, where $Head\ coach_t$ is the name of the head coach who was in charge of the club in the match that took place in round t.[9] Note that our analysis focuses on dismissals and does not consider cases of termination by mutual consent or voluntary quits by the old coach. Moreover, any match managed by a temporary caretaker manager is discarded from the analysis. From a careful inspection of the archives from the official websites of the league and individual clubs, as well as the two most-read national sports newspapers in Italy, *Corriere dello Sport-Stadio* and *La Gazzetta dello Sport*, we identified 157 cases during 2004/2005–2017/2018 to be included in the analysis.[10]

Given that each case of leadership change introduces simultaneous changes in managerial characteristics, investigating whether such characteristic changes can account for the effectiveness of replacement is a relevant issue in leadership succession. Therefore, we collected additional information related to the individual manager's characteristics from Transfermarkt (https://www.transfermarkt.com/). These include important indicators of leadership characteristics previously identified in sports economics. See, for instance, Bolton, Brunnermeier, and Veldkamp (2013), Bridgewater, Kahn, and Goodall (2011), Dawson and Dobson (2002) and Detotto, Paolini, and Tena (2018) for the managerial characteristic indicators related to professional sports. In a more general setting, the effect of CEO characteristics on corporate performance has also been studied (Kaplan, Klebanov, & Sorensen, 2012).

The first set of managerial characteristics is related to the individual manager's previous experience as a head coach in professional football leagues; experience in years (*Experience in years*), dummy variables that indicate whether: a manager had previously held a relevant role within *Serie A* (*Experience Serie A*), this is his first employment in the relevant role (*No previous experience*), a manager has previous experience in a top tier professional league abroad (*Experience abroad*). The second set of dummy variables is related to a manager's background as a professional player, which indicates whether: a manager is a former professional football player (*Former player*), a manager is a former player in *Serie A* (*Former player Serie A*), and a manager is a former defender or goalkeeper (*Former defender/goalkeeper*). The third set of indicators relates to a manager's association with the club. They indicate whether: the manager is a former vice coach of the club (*Former vice coach*), he is a former player of the club (*Former player club*), and the club is the last club with which he has been a player (*Last club as a player*). Another couple of dummy variables associated with recent employment status in the relevant role are considered. In particular, one takes a value equal to 1 if a manager was not employed in a relevant role in any club in the immediately preceding season, and 0 otherwise (*Absent last season*). The other indicator associated with recent activity indicates whether a manager was active or employed at any club participating in *Serie A* in the immediately preceding season (*Active Serie A last season*). The final set of variables are related to a manager's personal features: a manager's age in years (*Age in years*) and an indicator that takes a value equal to 1 if a manager is Italian, and 0 otherwise (*Italian nationality*).

Again, since each case of managerial succession results in changes in these managerial characteristics which could also affect post-succession performance, we take into account differences between the new and old managers. That is, for each characteristic variable h_t, we take the difference in the value of the variable between the manager in place at time t and the manager who had been in place at time $t-1$, i.e. $\Delta h = h_t - h_{t-1}$. Where a characteristic variable h_t is binary, as is the case for many of them, Δh is tertiary and takes values $\{-1, 0, 1\}$. Effectively, $\Delta h = 0$, where there was no managerial succession, or no difference between the new and old managers in the respective characteristic. Given this, Table 1 provides the summary statistics of the characteristic change variables for the 157 cases of managerial change considered in the analysis.

The Table provides some insight into the sort of changes made in managerial succession. In most cases, the value of Δh is equal or close to 0, implying that the new and old managers share a similar respective characteristic, hence suggesting a tendency towards clubs favouring like-with-like replacement. This may be because many clubs have a vision of what the ideal profile of a manager would be. However, there are also many cases of changes in the values of the characteristic variables. Therefore, in the subsequent analysis, we estimate the individual effect of changes in specific characteristics, other things being equal, on post-succession performance.

4.1.2. Variables related to treatment assignment

In order to estimate the propensity scores, a number of covariates which may affect the likelihood of treatment (i.e., head coach dismissal) are considered for inclusion in the treatment assignment model. These are identified in another strand of literature, for instance, Bryson, Buraimo, Farnell, and Simmons (2021) and references therein. The main cause of within-season managerial dismissal is related to the club's recent on-field performance, just like CEOs are often dismissed when firms are experiencing poor financial performance (Kato & Long, 2006; Hubbard, Christensen, & Graffin, 2017). We measure recent on-field performance by the average number of points earned over the last four matches (*Points last four matches*) and a dummy variable to indicate a loss in the most recent match (*Loss last match*). In addition, we include a dummy variable to indicate whether a defeat was at the club's home stadium (*Loss last match at*

[9] The league currently features 20 clubs, yielding the total number of matches played by an individual club in a given season of 38. Round t, therefore, corresponds to the t-th match in a particular season.

[10] During the relevant seasons, 15 cases of voluntary departures of head coaches were identified. For the same period of time, there were eight caretaker managers who were in charge during the transition between outgoing and incoming head coaches.

Table 1
Summary statistics of differences in managerial characteristics.

Variable	Difference between new and dismissed coaches (Δh)		
Binary indicators	**−1**	**0**	**1**
Former player	18 (11%)	120 (76.43%)	19 (12.1%)
Absent last season	18 (11%)	100 (63.69%)	39 (24.84%)
Former defender/goalkeeper	34 (22%)	99 (63.06%)	24 (15.29%)
Former vice coach	7 (4%)	136 (86.62%)	14 (8.92%)
Italian nationality	12 (8%)	132 (84.08%)	13 (8.28%)
Experience Serie A	25 (16%)	106 (67.52%)	26 (16.56%)
No previous experience	8 (5%)	133 (84.71%)	16 (10.19%)
Former player Serie A	36 (23%)	88 (56.05%)	33 (21.02%)
Former player club	18 (11%)	113 (71.97%)	26 (16.56%)
Last club as a player	7 (4%)	141 (89.81%)	9 (5.73%)
Experience abroad	25 (16%)	100 (63.69%)	32 (20.38%)
Active Serie A last season	41 (26%)	87 (55.41%)	29 (18.47%)
Continuous variables	**Min.**	**Mean**	**Max.**
Age in years	−26	1.080	29
Experience in years	−31	0.760	33

Notes: Table shows the summary statistics of managerial characteristics change variables for the 157 replacements included in the analysis. By construction, the difference variables for a binary characteristics indicator takes the value of −1, 0, and 1, where frequencies of each value together with the percentage of all the cases are reported. Whilst those for continuous variables are also continuous for which the maximum, mean, and minimum values are presented.

home), to account for the possibility that this event brings more pressure on a club than an away defeat. It has also been shown that performance relative to expectations matters. To take this into account, we include a measure of "surprise" accumulated over the relevant season (*Cumulative surprise*). Following van Ours and van Tuijl (2016), a surprise is measured by the deviation of actual points from expected points for each match, where expected points are obtained using the *ex-ante* probabilities of win, draw, and loss for each match based on the closing odds available from various bookmakers (https://www.football-data.co.uk/). We also consider the current league position relative to the final position in the previous season (*Relative standing*), which captures performance against subjective expectation by the fans. Furthermore, the current situation of a club is captured by two variables indicating whether a club is in the relegation zone (*Relegation zone*) and current position in the league (*Standing*), respectively. Whilst this study focuses on performance in the domestic league (*Serie A* in our case), it is possible that performance in other competitions could affect the prospect of a manager being dismissed. In particular, unfavourable outcomes, particularly critical ones, in other important competitions can impose extra pressure on the job security of a manager. To take this into account, we consider three binary variables indicating whether a club had been eliminated from UEFA Champions League (*Eliminated Champions League*), UEFA Europa League (*Eliminated Europa League*), or Coppa Italia (*Eliminated Coppa Italia*), between two *Serie A* matches t and $t-1$.

Additional variables considered in the treatment assignment model are an indicator of whether the club had already replaced a manager in the particular season (*Having dismissed this season*) and the number of days between two matches (*Days between matches*), which could potentially affect the decision of within-season managerial dismissals. Finally, as previous studies have shown, see Muehlheusser et al. (2016) for instance, within-season dismissals occur more frequently in mid-season. To capture this effect, we include round dummy variables in the treatment assignment model.

4.1.3. Outcome variables and additional control variables associated with the outcome

To measure club performance following treatment assignment, we construct outcome variables based on average points obtained in subsequent matches. For robustness, we obtain these values using up to five matches (*Points five matches*), ten matches (*Points ten matches*), and all of the remaining matches in the season (*Points rest of season*) or until the next managerial change, whichever occurs earlier.

In our outcome model, we include additional control variables that can affect post-treatment performance (i.e., whether a new coach is likely to earn points based on the focal match considered). First, a variable *Home advantage* controls for home advantage measured by the proportion of the matches that took place at the home stadium, out of the matches with which we measure the outcome variable. In addition, the ability level of the club (*Club ability*) and that of opponents (*Opponent ability*) are controlled by the ability indicator constructed in the following manner. First, we take a club's final position in the league table in the preceding season, reversing the order so that, for example, the top club was assigned the value 20 (and the bottom club would be assigned the value 1). The order is reversed to ensure that the variable increases with club ability as captured by its performance in the preceding season. In cases where a club had not played in the top division in the preceding season, it was assigned the value 1 (i.e. treated as having been equivalent to the bottom club in the top tier). We obtain these values for the final positions over the past four seasons, then take the weighted average with higher weights given to the more recent seasons for each club.[11] The variable *Opponent ability* is the average value of the ability indicator for the opponents in the subsequent matches with which the outcome is measured. We provide the correlation matrix of the variables used in the analysis in Appendix E of the supplementary material.

4.2. Methodology

We estimate the following outcome model to analyse the consequence of involuntary head coach replacements on y_{its}, our measure of the performance of club i, at round t in season s. Specifically, it is defined as:

$$y_{its} = \delta\, New\ coach_{its} + \gamma' X_{its} + \varepsilon_{its}, \tag{4}$$

where $New\ coach_{its} = 1$ if club i has replaced its manager prior to round t, and $New\ coach_{its} = 0$ otherwise; X_{its} is a vector of control variables. In particular, it includes variables related to managerial dismissal as well as variables associated with match outcomes.[12] A coefficient δ and a vector γ are parameters to be estimated. Finally, ε_{its} is a stochastic error component.

Our focus is to obtain the estimate of δ, which, if the treatment ($New\ coach_{its}$) were randomly allocated, should capture the Average Treatment Effect (ATE) of managerial change, if any. Model (4) is subsequently augmented to account for the various changes that may have been made with respect to the managerial characteristics of the head coach. Such possible changes are captured by a set of indicators defined as differences in managerial characteristic variables between replaced and appointed coaches, as explained in the previous section. Therefore, our extended model is specified as follows:

$$y_{its} = \delta\, New\ coach_{its} + \beta' \Delta H_{its} + \gamma' X_{its} + \varepsilon_{its}, \tag{5}$$

where ΔH_{its} is a vector of the managerial characteristic change variables, and β is its associated vector of parameters. Variables in ΔH_{its} take value zero when there was no managerial change prior to match t, or when no change was made with regards to the particular feature of the manager. Therefore, if parameters in vector β are significantly different from zero, this implies that differences in the characteristics

[11] More precisely, the weights given to the seasons $s-1, s-2, s-3, s-4$ are $0.5, 0.3, 0.15$, and 0.05, respectively, where s represents the current season. The idea is adapted from Dixon and Coles (1997), who suggest that a club's ability is better measured by recent performance with increasing weights on the more recent information.

[12] Including determinants of match outcome observed after the treatment is relevant in this setting as match score is also affected by home advantage and ability measures of both teams. All these variables can be considered exogenous as they occur in a quasi-random fashion.

between outgoing and incoming managers do matter for the successful implementation of managerial change.

An important concern in the estimation of models (4) and (5) is that head coach changes are not random events since they tend to occur more frequently with exceptionally low-performing clubs. Note that the inclusion of determinants of managerial dismissals allows us to control for different characteristics of treated and untreated teams. However, a simple OLS regression is not informative on whether these two groups are comparable in terms of their observable characteristics, threatening the causal interpretation of the estimation results. Under the assumption that we can observe the main determinants of managerial dismissal, PSA can be used to obtain counterfactuals that allow for a causal estimation.

A characteristic of our setting is that a head coach dismissal is a sporadic event, in the sense that 157 club-match observations out of 10,344 were followed by managerial replacement.[13] Thus, it is essential to find comparable counterfactuals in terms of observable variables for each treated observation. For this example, we choose PSW based on its simplicity and because it can include the whole sample in the estimation. Thus, since our treatment is binary, this implies that treated observations are weighted with the inverse of the probability of being treated, while control observations are given weights defined by the inverse of $(1 - \text{the probability of being treated})$. As a result, the distribution of propensity scores, i.e. the *ex-ante* probabilities of being treated, becomes similar between the treatment and control groups, as though the treatment were allocated randomly.

In our case, an observation is considered treated when a newly assigned manager is in charge at time *t* following the dismissal of a previous manager. Therefore, the likelihood of treatment assignment depends on the information related to performance that has been realised prior to time *t*. We estimate propensity scores by means of logistic regression. The model selection follows stepwise regression with a sequential replacement algorithm. The sequential replacement combines forward and backward selections, where the predictors proposed in Section 4.1 are iteratively added and removed until the lowest predictive error is achieved. We use the most common measure of predictive error, Akaike Information Criterion (AIC). See Bruce and Bruce (2017) for the details of stepwise regressions. Given the set of selected covariates, Z_{its}, we obtain the predicted values $\hat{p}_{its} = \Pr[\textit{New coach}_{its} = 1|Z_{its}]$, then the inverse propensity score weights are defined as follows:

$$w_{its} = \begin{cases} \frac{1}{\hat{p}_{its}}, & \text{if } \textit{New coach}_{its} = 1, \\ \frac{1}{1-\hat{p}_{its}}, & \text{if } \textit{New coach}_{its} = 0. \end{cases} \tag{6}$$

These weights may now be used in the weighted regression analysis to obtain the parameter estimates of models (4) and (5); see Guo and Fraser (2014) and Morgan and Todd (2008) for the use of inverse propensity score weights in the estimation of linear models. Moreover, a set of covariates selected in the treatment assignment model (Z_{its}) will be included in the outcome models according to the doubly robust estimation procedure. To appraise the relevance of such an approach, we report results using a “non-double robust” procedure in Appendix D of the supplementary material. In the following sections, we follow the steps described in Section 2 and Fig. 1 to estimate first the treatment assignment model, then the outcome models (4) and (5).

Step 1: Propensity score estimation

As described in Section 2, the initial step in PSW is to estimate the treatment assignment models. We estimate the logistic regression with stepwise selection, where we consider the set of covariates presented in the previous section in the initial step.[14] The set of covariates selected in the final model and estimation results are reported in Table 2.

The estimated coefficients of the selected covariates present expected signs,[15] being in line with previous findings. The probability of managerial change increases when a club has: lost the last match, performed poorly in the previous four games, and suffered negative surprising results during the season. In addition, the likelihood of turnover is higher when there are more days available between the last and current match. A recent elimination from the Europa League, as well as a threat of relegation, also contributes to a higher probability of managerial change.

Step 2: Obtaining PS weights

Having estimated the treatment assignment model, we will now have a closer look at the distribution of predicted values. First, the average predicted probabilities of treatment ($\hat{p}_{its} = \Pr[\textit{New coach}_{its} = 1|Z_{its}]$) are 0.0141 for those who did not change the manager (control group) and 0.0834 for those who actually did change their manager (treatment group). Estimates ranged from almost nil (4.395×10^{-6}) to 0.8191 for the former group and from 0.0016 to 0.5460 for the latter. Note that all the treated cases are contained within the common support, i.e. where the ranges of propensity scores for treated and control groups overlap. Using these predicted values, we then compute the weights according to the weighting function defined in (6).

Step 3: Balance diagnostic

We can now check whether the PSW can reduce the imbalancedness of the covariates included in the treatment assignment model. To do so, following Austin and Stuart (2015) and Morgan and Todd (2008), we compare the average value of absolute standardised mean differences (SMD) between the treated and control groups for each covariate. The standardised difference of the mean for a covariate z is calculated as:

$$\frac{|\bar{z}_{i,d_i=1} - \bar{z}_{i,d_i=0}|}{\sqrt{\frac{1}{2}\mathrm{Var}[z_{i,d_i=1}] + \frac{1}{2}\mathrm{Var}[z_{i,d_i=0}]}}, \tag{7}$$

where $\bar{z}_{i,d_i=1}$ and $\bar{z}_{i,d_i=0}$ are the means for those in treatment group ($d_i = 1$) and control group ($d_i = 0$), respectively, and $\mathrm{Var}[z_{i,d_i=1}]$ and $\mathrm{Var}[z_{i,d_i=0}]$ are the respective variances. The measure reflects the distance between the two groups in terms of the covariate that affects the treatment assignment. Table 3 presents these values for (1) raw sample, (2) weighted sample, and (3) weighted sample within common support. To balance treated and control groups, PSW should reduce differences between treated and control groups across all important covariates. Such differences are measured by the SMDs, and the average

[13] This low number of treated observations is also common in previous PSA papers (Bechtoldt et al., 2019; Li et al., 2021; Ong, 2021)

[14] Thoemmes and Ong (2016) indicate that PSW in the longitudinal case should repeat the process of weighting at every single point. The idea is to make treatment dependent only on information occurring before this decision, including also previous treatment decisions. In this example, we keep that spirit as model covariates only contain information that precedes treatment decisions. Moreover, a variable *Having dismissed this season* captures any previous decision to dismiss a manager in the same season.

[15] Note that these estimates could be dependent on the selection method employed. As noted in Section 2.2, other estimation or selection methods are available. For example, using Lasso results in a slightly different set of selected covariates and associated coefficients. In Appendix C of the supplementary material, we provide robustness exercises which compare the final estimations of ATE using different strategies. The estimated ATE remains similar under different propensity score estimation strategies.

Table 2
Stepwise regression[a] results for treatment assignment.

	Dependent variable:
	New coach
(Intercept)	-5.356^{***}(0.363)
Cumulative surprise	-0.253^{***}(0.025)
Days between matches	0.062^{***}(0.020)
Points last four matches	-0.800^{***}(0.190)
Loss last match	1.424^{***}(0.252)
Relegation zone	0.393^{**}(0.186)
Eliminated Europa League	1.309^{*}(0.769)
Observations	10,344
Log Likelihood	−615.719
Akaike Inf. Crit.	1,245.438

Notes: [a]The stepwise regression with the lowest AIC as a stopping criterion. *p<0.1; **p<0.05; $^{*\,*\,*}$p<0.01. Robust standard errors are in parentheses.

values of SMDs across all the important covariates are 0.796, 0.312 and 0.127 in the raw sample, weighted sample, and weighted sample with common support, respectively. They suggest significant overall reductions in imbalancedness due to weighting and a further improvement in balance within common support. Furthermore, including observations outside of the common support would mean that our estimation of ATE partly relies on observations for which counterpart observations are not available. Therefore, we use the sample within the common support to estimate the consequences of head coach turnover in the following subsection.

Step 4: Estimation of treatment effects

The final step is to estimate the treatment effects by applying the defined weights through weighted regression analysis. Before we proceed, however, we briefly discuss the possible consequences of not addressing the imbalancedness detected in the previous steps. In particular, the differences in preceding performances between the treated and control groups are large, implying that involuntary managerial changes are not random events. However, no theory provides a clear indication of how ignoring such differences can affect conclusions on the impact of replacing a manager. On the one hand, one can argue that poorly-performing teams may revert to their mean performance levels regardless of whether they replace their head coaches. On the other hand, it is also plausible to assume that some poorly-performing teams are more likely to carry on with this negative inertia due to persistent issues, such as long-term injuries or conflicts among players, even if they replace their head coach. Fig. 2 plots the average values of performance in the post-treatment periods (between treatment assignment and the end of the respective season) for treated and control groups, at a given level of the club's ability in the raw sample. The initial look of the Figure suggests that performance is increasing in a club's ability; however, no *prima facie* difference between treatment and control groups is evident in the raw sample.

The OLS estimates of model (4) suggest some weak evidence of detrimental effects of managerial change. The details of OLS estimation are available in Appendix D of the supplementary material. Of course, this approach is not robust to the potential selection bias discussed above since it does not focus on comparable treated and control groups in terms of observable characteristics.

Now we revert to the estimation of our outcome models (4) and (5), using PSW. The estimation results for these models are shown in Table 4, where outcome variables are the average points obtained in the post-treatment matches, where we include up to 5 matches, 10 matches, and the rest of the season.

The estimates for model (4)[16] are reported in columns (1), (3), and (5), for the respective outcome variables. No significant treatment effects at the 10% significance level are detected in the short run (first five matches). Still, a positive and significant impact at the 5% significance level is evident once a longer run of post-treatment matches is considered.

Columns (2), (4), and (6) in Table 4 present the estimated parameters for our extended model (5), which includes the additional variables capturing differences in characteristics between the new and outgoing managers. Including these variables does not affect the sign of the binary treatment effect (*New coach*) in the corresponding baseline models (1), (3), and (5), respectively, but its size is smaller. The results suggest that the changes in particular characteristics of managers affect post-treatment performance. For instance, when a new manager was absent (not employed as a head coach elsewhere) in the previous season, this tends to have a positive impact on post-succession outcomes. On the other hand, older replacement managers tend to achieve a negative treatment effect. The variables that capture the changes associated with experiences, such as experience in years, experience abroad, experience in *Serie A*, and no previous experience, do not show significant effects to explain the post-succession performance at the 10% significance level. Similarly, having been employed at a *Serie A* club in the immediately preceding season is not a significant variable at the 10% significance level.

A new manager's background as a professional player relative to that of a dismissed manager, in general, does not have a significant impact at the conventional significance levels, whilst a positive outcome is expected if a manager played a more defensive role as a player. However, when a new manager is a previous *Serie A* player and a dismissed one is not, the succession tends to have a negative effect holding other things constant. A speculative explanation for this is that becoming a manager in a new market (where they did not participate as a player) indicates desirable managerial skills. The positive and significant coefficient of *Last club as a player* implies that a manager with a stronger association with the club (one who finished his playing career at the club) can positively influence post-succession performance whilst merely being a former player of the club (*Former player club*) has no significant effect at the 10% significance level. However, replacing a manager with a former vice coach at the club tends to have a detrimental effect, particularly in the short term. Finally, changes in nationality, i.e. being Italian, do not show any significant impact at the conventional levels.

In all the models, the coefficient estimates on all the control variables have expected signs; the percentage of home matches and club ability both have a significant positive effect on match outcomes, and a club's performance is negatively correlated with the average ability of their opponent clubs.

Some robustness exercises are reported in Appendix C in the supplementary material. These exercises address the relevance of the PS specification and the uncertainty that stems from using a two-step procedure in the final ATE estimation. Our estimated ATEs remain similar under each of the several alternative strategies explored in the appendix.

4.3. Extension: endogeneity of similarity in coach characteristics

In the last step of our previous analysis, we did not account for the potential endogeneity of changes in managerial characteristics in the causal estimation. This decision could be justified in this context because the scope for selecting each dimension of managerial characteristics is limited given the limited time and candidates available in the within-season setting. Nevertheless, we can consider the possibility of a club endogenously choosing a similar or dissimilar replacement in terms of overall characteristics. Therefore, in this more advanced

[16] Following Ridgeway et al. (2021), the estimations of outcome models are obtained using the "svyglm" function in R, which is commonly used for survey sample analysis and automatically produces robust standard errors.

Table 3
Covariate balance table (before/after weighting, all/common support).

	Raw		Weighted		Weighted (CS)	
Covariate	SMD	P-value	SMD	P-value	SMD	P-value
Cumulative surprise	1.265	0.000	0.658	0.000	0.199	0.199
Days between matches	0.284	0.004	0.107	0.271	0.041	0.654
Eliminated Europa League	0.074	0.500	0.064	0.000	0.077	0.000
Points last four matches	1.107	0.000	0.622	0.000	0.120	0.350
Loss last match	1.044	0.000	0.133	0.402	0.268	0.082
Relegation zone	1	0.000	0.290	0.070	0.058	0.671
Mean SMD	0.796		0.312		0.127	
N (Treated)	157		157		157	
N (Control)	10187		10187		6218	

Notes: Table reports the absolute values of standardised mean differences (SMD) between the treatment and control groups before and after weighting.

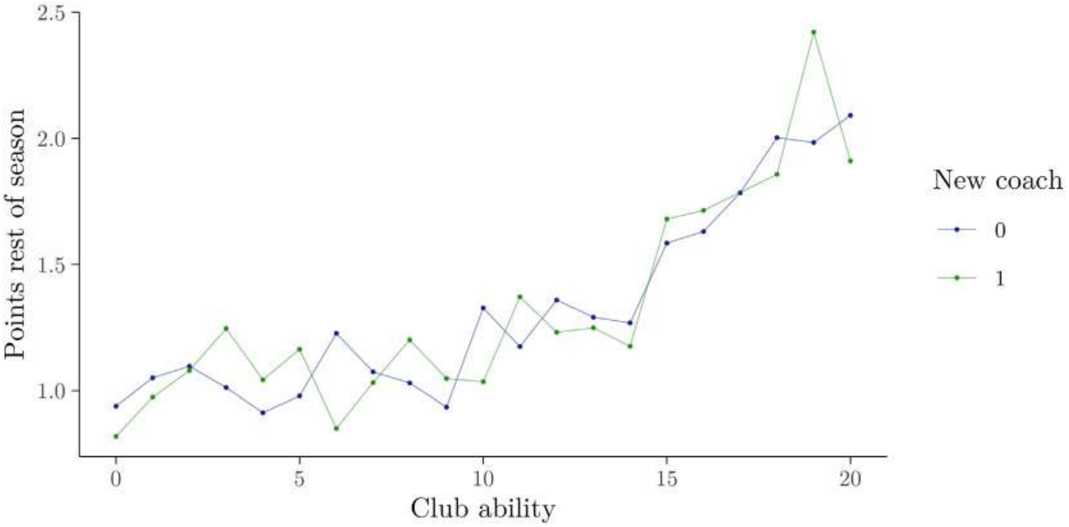

Fig. 2. Mean value of outcome variable (*Points rest of season*) for different levels of ability and treatment group. The x-axis represents the ability of the club (*Club ability*), computed based on the weighted average of the final league positions in the preceding four seasons, with the value 1 being the lowest ability and 20 being the highest. The y-axis measures the mean values of an outcome variable (*Points rest of season*), the average points obtained following assignment or non-assignment of the treatment for each treatment group (*New coach* = 0 and *New coach* = 1).

example, our purpose is to illustrate how to modify the analysis to consider a multi-level treatment, where a club can decide further whether the replacement should be similar or dissimilar to the dismissed manager. As we discussed in Section 3, dealing with a non-binary treatment is a common problem in empirical research.

To define the similarity between the new and dismissed manager, we cluster managers using the characteristic variables included in the previous analysis. In particular, we employ the Partitioning Around Medoid (PAM) algorithm[17] to group the managers into clusters based on the similarities in terms of their characteristics. We then define the treatment as "similar" if dismissed and appointed managers are in the same cluster, and "dissimilar," if they are in distinct clusters. More formally, we create an additional dummy variable *Dissimilar coach*, where *Dissimilar coach* = 0 if the new coach is "similar" to the dismissed according to the above definition, and *Dissimilar coach* = 1 if the new coach is "dissimilar." Based on this definition, we identify 112 cases out of 157 cases of the replacements as "dissimilar" changes and 45 cases as "similar" changes. To incorporate this additional layer into the decision problem, we consider a nested logistic regression to obtain the probabilities of no change, similar change, and dissimilar change. The first nest models the decision regarding whether to replace a manager (*New coach* = 1) or not (*New coach* = 0), as considered in the previous section. The model of the second nest estimates the probability of a dissimilar replacement (*Dissimilar coach* = 1), within the treated observations (*New coach* = 1). A graphical representation of the nested logit model is given in Fig. 3, which also illustrates the three treatment types and associated probabilities.

The procedure, akin to Step 1 in the previous section, can be applied to estimate the probability of dissimilar change, i.e. $\hat{\Pr}[\textit{Dissimilar coach} = 1|Z]$. The estimation results are quite different from those in Table 2, where only a few covariates were selected: *Relegation zone*, *Days between matches*, and *Standing*. Underlying imbalancedness is also less severe. The SMDs of the selected covariates between the dissimilar and similar changes are not significant at the 5% significance level in the raw sample, and the average value of the absolute SMDs is.217. Nevertheless, a significant reduction in the SMDs is achieved between the similar and dissimilar changes by applying the weights defined by the inverse of respective propensity scores; the average value of the absolute SMDs is.026 in the weighted sample.

Based on this, we extend our previous model to assess the effectiveness of the three possible treatments, (1) no change, (2) similar change, and (3) dissimilar change. As explained in Wooldridge (2010), regression adjustment in the multiple treatment case is an obvious extension of the case where treatment is binary. Therefore,

[17] This method for clustering is suitable for our context, where a mix of continuous and categorical variables is to be considered. The algorithm identifies the optimal number of clusters based on "silhouette widths", a measure of relative similarity to the members in the same group compared to those in the other group. See der Laan, Pollard, and Bryan (2003) for the details.

Table 4
Double robust estimates of outcome models.

	Dependent variable:					
	Points five matches		Points ten matches		Points rest of season	
	(1)	(2)	(3)	(4)	(5)	(6)
New coach	0.139	0.102	0.184*	0.137**	0.180**	0.117**
	(0.104)	(0.063)	(0.097)	(0.061)	(0.090)	(0.056)
Former player		0.120		0.048		0.049
		(0.105)		(0.110)		(0.103)
Absent last season		0.163		0.256**		0.288***
		(0.107)		(0.102)		(0.097)
Age in years		−0.009		−0.019*		−0.020**
		(0.011)		(0.011)		(0.010)
Experience in years		−0.007		0.009		0.013
		(0.010)		(0.010)		(0.009)
Former defender/goalkeeper		0.164*		0.242***		0.186**
		(0.094)		(0.091)		(0.082)
Former vice coach		−0.454***		−0.170		−0.210
		(0.174)		(0.188)		(0.182)
Italian nationality		0.198		0.171		0.068
		(0.124)		(0.154)		(0.143)
Experience Serie A		0.006		−0.035		−0.126
		(0.117)		(0.117)		(0.114)
No previous experience		0.055		−0.085		−0.109
		(0.198)		(0.165)		(0.152)
Former player Serie A		−0.365***		−0.204*		−0.193**
		(0.099)		(0.105)		(0.090)
Former player club		0.033		−0.065		0.026
		(0.150)		(0.155)		(0.136)
Last club as a player		0.523**		0.665***		0.529***
		(0.232)		(0.216)		(0.191)
Experience abroad		−0.110		−0.069		−0.061
		(0.105)		(0.104)		(0.096)
Active Serie A last season		0.094		0.075		−0.012
		(0.086)		(0.098)		(0.091)
Home advantage	1.101***	0.849***	1.057***	0.773***	1.029***	0.802***
	(0.230)	(0.139)	(0.261)	(0.149)	(0.298)	(0.151)
Club ability	0.057***	0.038***	0.059***	0.043***	0.061***	0.043***
	(0.015)	(0.004)	(0.013)	(0.004)	(0.013)	(0.004)
Opponent club ability	−0.032**	−0.040***	−0.039*	−0.046***	−0.031	−0.035**
	(0.015)	(0.010)	(0.022)	(0.014)	(0.023)	(0.015)
Constant	0.638***	0.807***	0.797***	0.949***	0.724***	0.809***
	(0.155)	(0.117)	(0.195)	(0.131)	(0.229)	(0.152)
Observations	6,375	6,375	6,375	6,375	6,375	6,375
Log Likelihood	−7,998.567	−7,261.832	−7,361.029	−6,685.688	−7,036.235	−6,288.864
Akaike Inf. Crit.	16,019.130	14,573.670	14,744.060	13,421.380	14,094.470	12,627.730

Notes: *p<0.1; **p<0.05; $^{*\,*\,*}$p<0.01. Robust standard errors are in parentheses. The estimation includes the covariates selected for the propensity score estimation as control variables. However, the estimated coefficients associated with these controls are not reported.

we weight the sample with the inverse of the *ex-ante* probability of actual treatment status, as depicted in Fig. 3. Then, we estimate the outcome model (4) with an additional treatment variable *Dissimilar coach*, together with the control variables associated with the outcome and the covariates selected in the treatment assignment model. The results are reported in Table 5. The estimated coefficients of *New coach* and *Dissimilar coach* indicate that replacement with a similar manager has no statistically significant effect at the 10% significance level, whilst the appointment of a new manager who has a different profile than the old is associated with an improvement in the following five and ten matches at the 10% and 5% significance levels, respectively.

4.4. Discussion

Although our estimation exercise has mainly a didactical purpose, some results are noteworthy. Estimation results reported in Tables 4 indicate that the replacement of a head coach has, on average, a positive impact on subsequent performance. These results are obtained using PSW with double-robust estimation. Without using such a method, the results would not be as trustworthy since there are fundamental differences between the treated and control group, as is clear from Table 2. Moreover, we show how to extend the standard binary analysis by decomposing a head coach replacement into changes in different managerial attributes between the old and the new manager in a way that we can assess their separate impact.

The example shows that taking into account the differences between the new and dismissed coaches does provide further insights into the effectiveness of leadership change. For example, when a new manager has a stronger association with the club, indicated by the manager having finished his playing career at the club, this can positively influence post-succession performance. The negative (and significant in the short term at conventional levels) coefficients on *Former vice coach* imply that internal succession is expected to worsen a club's performance. This can be partly explained by the view that the internal succession may involve more minor strategic change due to cognitive and psychological attachment to the existing strategy (Farah et al., 2020). The analysis also shows that appointing a new head coach who had not been in employment as a coach in the preceding season could be effective. Recent absence could be a desirable managerial characteristic since engaging in activities outside coaching and reflecting on their working methods may help them adopt a broader perspective.

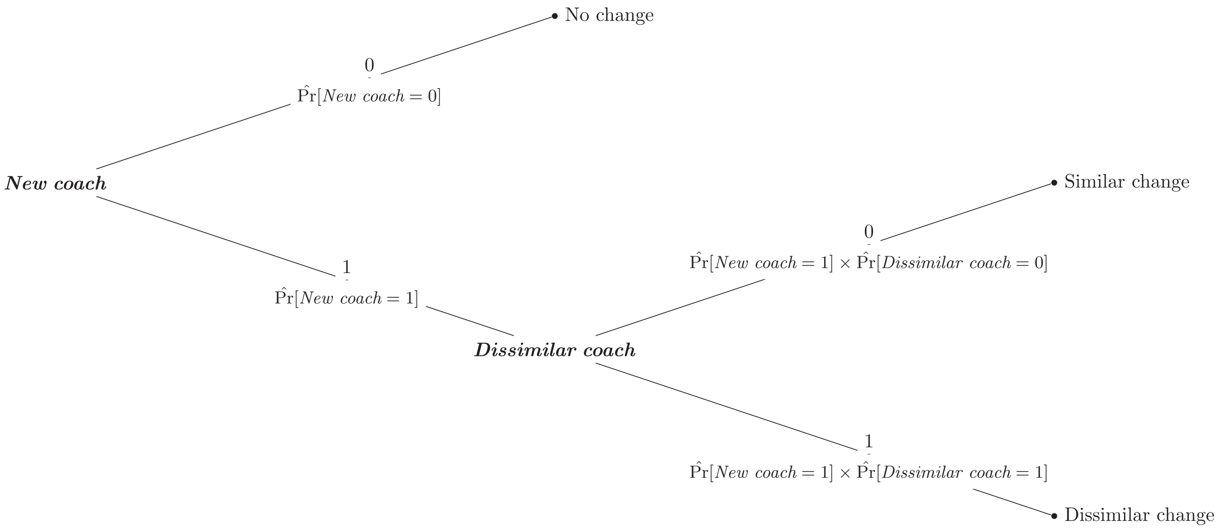
No change
0
$\hat{\Pr}[New\ coach = 0]$
New coach
Similar change
1
0
$\hat{\Pr}[New\ coach = 1] \times \hat{\Pr}[Dissimilar\ coach = 0]$
$\hat{\Pr}[New\ coach = 1]$
Dissimilar coach
1
$\hat{\Pr}[New\ coach = 1] \times \hat{\Pr}[Dissimilar\ coach = 1]$
Dissimilar change

Fig. 3. Nested logit model The Figure illustrates the nested logit model, where the first nest classifies the cases into *New coach* $= 0$ or *New coach* $= 1$, and the second nest further classifies cases with *New coach* $= 1$ into *Dissimilar coach* $= 0$ or *Dissimilar coach* $= 1$, resulting in the three possible outcomes (No change, similar change, and dissimilar change). Corresponding probabilities of each outcome are obtained using the predicted values resulting from the estimation of logistic regression of each nest.

Table 5
Double robust estimates of outcome model with dissimilar treatment.

	Dependent variable:		
	Points 5 matches	Points 10 matches	Points rest of season
	(1)	(2)	(3)
New coach	−0.084	0.004	0.060
	(0.121)	(0.075)	(0.066)
Dissimilar new coach	0.247*	0.195**	0.117
	(0.147)	(0.090)	(0.084)
Home advantage	1.080***	0.985***	0.894***
	(0.253)	(0.249)	(0.259)
Club ability	0.036***	0.028***	0.030***
	(0.010)	(0.007)	(0.007)
Opponent club ability	−0.026*	−0.028	−0.014
	(0.014)	(0.020)	(0.021)
Constant	1.286***	1.982***	1.873***
	(0.415)	(0.211)	(0.242)
Observations	6,375	6,375	6,375
Log Likelihood	−8,557.107	−7,595.530	−7,228.201
Akaike Inf. Crit.	17,140.210	15,217.060	14,482.400

Notes: *p<0.1; **p<0.05; $^{* * *}$p<0.01. Robust standard errors are in parentheses. The estimation includes the covariates selected for the propensity score estimation as control variables. However, the estimated coefficients associated with these controls are not reported.

Some potential limitations of our empirical example are the following. First, as is the case in all analyses based on PSA, our results are causally interpretable only if there are no unmeasured confounders. Second, our example does not fully utilise the panel structure of the data, for instance, we do not consider fixed effects in our models. The extension of PSA for panel regression models with fixed effects is, however, not simple. For example, the panel should allow for within-cluster comparisons of treatment and control individuals with similar covariates. Moreover, individual units and treatment effects could change over time, so they cannot be considered fixed. Arkhangelsky and Imbens (2018) and Imai and Kim (2019) have recently discussed the use of fixed-effect models in causal analysis. Using panel samples that allow for applying these methods in leadership is a relevant avenue for future research.

5. Lessons, limitations and implications for future research

Randomised control trials could be unfeasible when empirical research involves the analysis of behaviour, emotions or decisions. In this paper, we explain how to conduct a PSW and discuss the implementation of this approach in recent papers in the management, applied psychology, and leadership literature. PSW is illustrated with an advanced tutorial case that estimates the causes and consequences of head coach changes in Italian football. The example presented in this paper also illustrates how to extend the analysis to estimate how different types of managerial dismissal affect post-succession performance. The tutorial approach is conceptually and methodologically advanced yet is simple to implement as it only requires the use of propensity scoring in weighted regression. We also demonstrate how to extend the analysis by considering managerial turnover as changes in multiple managerial attributes and estimating separate effects from such changes.

Although the example presented is specific to the sports industry, the particular nature of professional sports facilitates tackling internal validity concerns typically present in causal analysis. Giambatista et al. (2005) noted that while it is unclear whether results for specific sectors could be generalised elsewhere, non-sports contexts in the literature are also concentrated in very specialised settings such as manufacturing enterprises. Therefore, given the advantages of transparency in organisational objectives and measures of performance, they recommend researchers continue exploiting sports data to investigate issues around managerial succession. We hope this tutorial contributes to incentivising the use of sports data in future management and leadership research, which has become more recognised in the field (Fonti et al., 2022).

Three future lines of research can be proposed based on this study. The first possibility concerns considering more advanced methodologies such as machine learning (Doornenbal, Spisak, & van der Laken, 2021) for causal analysis. PS estimated with a machine learning approach can be easily integrated into the estimation process described in the tutorial without the need to wait for statistical packages that include the new methods in the matching algorithm. A second possibility is to explore further how managerial change is operationalised. Thus, future research could study, for example, how changes in head coach characteristics interact with organisational and environmental attributes or extend the set of managerial characteristics to include specific leadership behaviours, like charismatic signalling (e.g., Tur, Harstad, & Antonakis (2021)). A third possible future line of research is to use PSA to explore critical questions in the leadership literature, such as, for example, the effect of awards on performance or the impact of different types of leader decisions. The joint consideration of PSA and sports data seems, in principle, a promising avenue for future research.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Research data

Mendeley Data link is https://data.mendeley.com/datasets/xgkp4rkyf6.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to the editor Sirio Lonati, the method editor, and three anonymous referees for their guidance and constructive feedback. We would also like to thank David Forrest, Oliver De Groot, Benjamin Holmes, and Babatunde Buraimo, for useful comments and suggestions.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.leaqua.2023.101678.

References

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*(6), 1086–1120.

Arkhangelsky, D., and Imbens, G. (2018). The role of the propensity score in fixed effect models. *National Bureau of Economic Research*, No. w24814.

Audas, R., Dobson, S., & Goddard, J. (2002). The impact of managerial change on team performance in professional sports. *Journal of Economics and Business, 54*(6), 633–650.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399–424.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine, 34*(28), 3661–3679.

Bechtoldt, M. N., Bannier, C. E., & Rock, B. (2019). The glass cliff myth? – Evidence from Germany and the U.K. *The Leadership Quarterly, 30*(3), 273–297.

Berns, K. V., & Klarner, P. (2017). A Review of the CEO succession Literature and a Future Research Program. *Academy of Management Perspectives, 31*(2), 83–108.

Boivie, S., Graffin, S. D., Oliver, A. G., & Withers, M. C. (2016). Come Aboard! Exploring the Effects of Directorships in the Executive Labor Market. *Academy of Management Journal, 59*(5), 1681–1706.

Bolton, P., Brunnermeier, M. K., & Veldkamp, L. (2013). Leadership, coordination, and corporate culture. *Review of Economic Studies, 80*(2), 512–537.

Bridgewater, S., Kahn, L. M., & Goodall, A. H. (2011). Substitution and complementarity between managers and subordinates: Evidence from British football. *Labour Economics, 18*(3), 275–286.

Bruce, P., & Bruce, A. (2017). *Practical statistics for data scientists*. California, USA: O'Reilly.

Bryson, A., Buraimo, B., Farnell, A., & Simmons, R. (2021). Time To Go? Head Coach Quits and Dismissals in Professional Football. *De Economist, 169*(1), 81–105.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72.

Chen, G. (2015). Initial compensation of new CEOs hired in turnaround situations. *Strategic Management Journal, 36*(12), 1895–1917.

Connelly, B. S., Sackett, P. R., & Waters, S. D. (2013). Balancing treatment and control groups in quasi-experiments: An introduction to propensity scoring. *Personnel Psychology, 66*(2), 407–442.

Cook, T., & Campbell, D. (1976). The design and conduct of true experiments and quasi-experiments in field settings. In M. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology* (pp. 223–326). Chicago: Rand McNally.

Crano, W. D., Brewer, M. B., & Lac, A. (2014). *Principles and methods of social research* (3rd ed.). New York: Routledge.

Dawson, P., & Dobson, S. (2002). Managerial efficiency and human capital: An application to English association football. *Managerial and Decision Economics, 23*(8), 471–486.

de Jong, A., & Naumovska, I. (2016). A note on event studies in finance and management research. *Review of Finance, 20*(4), 1659–1672.

der Laan, M. J. V., Pollard, K. S., & Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation, 73*(8), 575–584.

Detotto, C., Paolini, D., & Tena, J. D. (2018). Do managerial skills matter? An analysis of the impact of managerial features on performance for Italian football. *Journal of the Operational Research Society, 69*(2), 270–282.

Devarakonda, S. V., & Reuer, J. J. (2018). Knowledge sharing and safeguarding in R&D collaborations: The role of steering committees in biotechnology alliances. *Strategic Management Journal, 39*(7), 1912–1934.

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society. Series C: Applied Statistics, 46*(2), 265–280.

Doornenbal, B. M., Spisak, B. R., & van der Laken, P. A. (2021). Opening the black box: Uncovering the leader trait paradigm through machine learning. *The Leadership Quarterly*, 101515 (In press).

Emsley, R., Lunt, M., & Dunn, G. (2008). Implementing double-robust estimators of causal effects.The. *Stata Journal, 8*(3), 334–353.

Farah, B., Elias, R., De Clercy, C., & Rowe, G. (2020). Leadership succession in different types of organizations: What business and political successions may learn from each other. *The Leadership Quarterly, 31*(1), 101289.

Fest, S., Kvaløy, O., Nieken, P., & Schöttner, A. (2021). How (not) to motivate online workers: Two controlled field experiments on leadership in the gig economy. *The Leadership Quarterly, 32*(6), 101514.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Fong, C., Hazlettand, C., & Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Annals of Applied Statistics, 12*(1), 156–177.

Fonti, F., Ross, J. M., & Aversa, P. (2022). Using sports data to advance management research: A review and a guide for future studies. *Journal of Management*.

Freedman, D. A., & Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review, 32*(4), 392–409.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology, 173*(7), 761–767.

Giambatista, R. C., Rowe, W. G., & Riaz, S. (2005). Nothing succeeds like succession: A critical review of leader succession literature since 1994. *The Leadership Quarterly, 16*(6), 963–991.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*(4), 405–420.

Guo, S. G., & Fraser, M. W. (2014). *Propensity score analysis: Statistical Methods and Applications* (2nd ed.). Thousand Oaks, CA: SAGE Publications Inc.

Gupta, V. K., Han, S., Mortal, S. C., Silveri, S., & Turban, D. B. (2017). Do women CEOs face greater threat of shareholder activism compared to male CEOs? A role congruity perspective. *Journal of Applied Psychology, 103*(2), 228–236.

Gupta, V. K., Mortal, S., Chakrabarty, B., Guo, X., & Turban, D. B. (2020). CFO gender and financial statement irregularities. *Academy of Management Journal, 63*(3), 802–831.

Hamilton, B. H., & Nickerson, J. A. (2003). Correcting for endogeneity in strategic management research. *Strategic Organization, 1*(1), 51–78.

Heinze, G., & Jüni, P. (2011). An overview of the objectives of and the approaches to propensity score analyses. *European Heart Journal, 32*, 1704–1708.

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 73–84). John Wiley & Sons Ltd.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Hopp, C., & Pruschak, G. (2020). Is there such a thing as leadership skill? – A replication and extension of the relationship between high school leadership positions and later-life earnings. *The Leadership Quarterly*, 101475 (In press).

Hubbard, T. D., Christensen, D. M., & Graffin, S. D. (2017). Higher highs and lower lows: The role of corporate social responsibility in ceo dismissal. *Strategic Management Journal, 38*(11), 2255–2265.

Hughes, M., Hughes, P., Mellahi, K., & Guermat, C. (2010). Short-term versus Long-term Impact of Managers: Evidence from the Football Industry. *British Journal of Management, 21*(2), 571–589.

Imai, K., & Kim, I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science, 63*, 467–490.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87*(3), 706–710.

Kaplan, S. N., Klebanov, M. M., & Sorensen, M. (2012). Which CEO Characteristics and Abilities Matter? *Journal of Finance, 67*(3), 973–1007.

Kato, T., & Long, C. (2006). Ceo turnover, firm performance, and enterprise reform in china: Evidence from micro data. *Journal of Comparative Economics, 34*(4), 796–817.

Kiss, A. N., Cortes, A. F., & Herrmann, P. (2021). CEO proactiveness, innovation, and firm performance. *The Leadership Quarterly*, 101545 (In press).

Larcker, D. F., & Rusticus, T. O. (2010). On the use of instrumental variables in accounting research. *Journal of Accounting and Economics, 49*(3), 186–205.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3).

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLoS ONE, 6*(3).

Li, M. (2013). Using the propensity score method to estimate causal effects: A review and practical guide. *Organizational Research Methods, 16*(2), 188–226.

Li, W. D., Li, S., Feng, J. J., Wang, M., Zhang, H., Frese, M., & Wu, C. H. (2021). Can becoming a leader change your personality? An investigation with two longitudinal studies from a role-based perspective. *Journal of Applied Psychology, 106*(6), 882–901.

Love, E. G., Lim, J., & Bednar, M. K. (2017). The face of the firm: The influence of CEOs on corporate reputation. *Academy of Management Journal, 60*(4), 1462–1481.

Luo, X. R., Zhang, J., & Marquis, C. (2016). Mobilization in the internet age: Internet activism and corporate response. *Academy of Management Journal, 59*(6), 2045–2068.

Morgan, S. L., & Todd, J. J. (2008). 6. A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology, 38*(1), 231–282.

Muehlheusser, G., Schneemann, S., & Sliwka, D. (2016). The impact of managerial change on performance: The role of team heterogeneity. *Economic Inquiry, 54*(2), 1128–1149.

Ong, W. J. (2021). Gender-contingent effects of leadership on loneliness. *Journal of Applied Psychology*. Advance online publication.

Podsakoff, P. M., & Podsakoff, N. P. (2019). Experimental designs in management and leadership research: Strengths, limitations, and recommendations for improving publishability. *The Leadership Quarterly, 30*(1), 11–33.

Raad, H., Cornelius, V., Chan, S., Williamson, E., & Cro, S. (2020). An evaluation of inverse probability weighting using the propensity score for baseline covariate adjustment in smaller population randomised controlled trials with a continuous outcome. *BMC Medical Research Methodology, 20*(1), 1–12.

Ridgeway, G., McCaffrey, D., Morral, A., Cefalu, M., Burgette, L., Pane, J., and Griffin, B. A. (2021). Toolkit for weighting and analysis of nonequivalent groups: A guide to the twang package. *vignette*, July, 26.

Rocha, V., & Van Praag, M. (2020). Mind the gap: The role of gender in entrepreneurial career choice and social influence by founders. *Strategic Management Journal, 41*(5), 841–866.

Rockey, J. C., Smith, H. M., & Flowe, H. D. (2021). Dirty looks: Politicians' appearance and unethical behaviour. *The Leadership Quarterly, 33*(2), 101561.

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.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39*(1), 33–38.

Rowe, W. G., Cannella, A. A., Rankin, D., & Gorman, D. (2005). Leader succession and organizational performance: Integrating the common-sense, ritual scapegoating, and vicious-circle succession theories. *The Leadership Quarterly, 16*(2), 197–219.

Schmidt, J. A., & Pohler, D. M. (2018). Making stronger causal inferences: Accounting for selection bias in associations between high performance work systems, leadership, and employee and customer satisfaction. *Journal of Applied Psychology, 103*(9), 1001–1018.

Semadeni, M., Withers, M. C., & Trevis Certo, S. (2014). The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations. *Strategic Management Journal, 35*(7), 1070–1079.

Shang, G., & Rönkkö, M. (2022). Empirical research methods department: Mission, learnings, and future plans. *Journal of Operations Management, 68*(2), 114–129.

Shi, W., Zhang, Y., & Hoskisson, R. E. (2019). Examination of CEO–CFO social interaction through language style matching: Outcomes for the CFO and the organization. *Academy of Management Journal, 62*(2), 383–414.

Stone, C. A., & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research and Evaluation, 18*(13), 1–12.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.

Sy, T., Horton, C., & Riggio, R. (2018). Charismatic leadership: Eliciting and channeling follower emotions. *The Leadership Quarterly, 29*(1), 58–69.

Tena, J. D., & Forrest, D. (2007). Within-season dismissal of football coaches: Statistical analysis of causes and consequences. *European Journal of Operational Research, 181*(1), 362–373.

Thoemmes, F., & Ong, A. D. (2016). A primer on inverse probability of treatment weighting and marginal structural models. *Emerging Adulthood, 4*(1), 40–59.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*(1), 90–118.

Tur, B., Harstad, J., & Antonakis, J. (2021). Effect of charismatic signaling in social media settings: Evidence from TED and Twitter. *The Leadership Quarterly, 101476*.

van Ours, J. C., & van Tuijl, M. A. (2016). In-season head-coach dismissals and the performance of professional football teams. *Economic Inquiry, 54*(1), 591–604.

Vitanova, I. (2021). Nurturing overconfidence: The relationship between leader power, overconfidence and firm performance. *The Leadership Quarterly, 32*(4), 101342.

Wofford, J. C. (1999). Laboratory research on charismatic leadership: Fruitful or futile? *The Leadership Quarterly, 10*(4), 523–529.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. London: MIT Press.

Zhang, Z., Zhang, B., & Jia, M. (2021). The military imprint: The effect of executives' military experience on firm pollution and environmental innovation. *The Leadership Quarterly, 33*(2), 101562.

Zheng, W., Singh, K., & Chung, C. N. (2017). Ties to unbind: Political ties and firm sell-offs during institutional transition. *Journal of Management, 43*(7), 2005–2036.