JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **23**: 305–327 (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/jae.998

FROM TEMPORARY HELP JOBS TO PERMANENT EMPLOYMENT: WHAT CAN WE LEARN FROM MATCHING ESTIMATORS AND THEIR SENSITIVITY?

ANDREA ICHINO,[a†] FABRIZIA MEALLI[b] AND TOMMASO NANNICINI[c*]

[a] *Department of Economics, University of Bologna, Bologna, Italy*
[b] *Department of Statistics, University of Florence, Florence, Italy*
[c] *Department of Economics, Universidad Carlos III de Madrid, Madrid, Spain*

SUMMARY

The diffusion of temporary work agency (TWA) jobs has led to a harsh policy debate and ambiguous empirical evidence. Results for the USA, based on quasi-experimental evidence, suggest that a TWA assignment decreases the probability of finding a stable job, while results for Europe, based on the conditional independence assumption (CIA), typically reach opposite conclusions. Using data for two Italian regions, we rely on a matching estimator to show that TWA assignments can be an effective springboard to permanent employment. We also propose a simulation-based sensitivity analysis, which highlights that only for one of these two regions are our results robust to specific failures of the CIA. We conclude that European studies based on the CIA should not be automatically discarded, but should be put under the scrutiny of a sensitivity analysis like the one we propose. Copyright © 2008 John Wiley & Sons, Ltd.

Received 7 February 2005; Revised 4 April 2007

1. INTRODUCTION

The growing share of temporary employment in many European countries has raised concerns over the risk of an undesirable labour market 'segmentation'. Several studies have indicated the existence of a gap in the working conditions of permanent and temporary employees, particularly in terms of wages and working rights.[1] Triggered by this gap, public opinion and policy makers have stressed the importance of finding 'an appropriate balance between flexibility and security' (European Commission, 2003). While such a balance may not be possible in a cross-sectional sense, it may become possible in an intertemporal sense if temporary jobs are an effective springboard toward permanent employment, as opposed to a trap of endless precariousness.

From a theoretical point of view, there are two broad reasons why temporary employment could offer a springboard to a stable job. First, more able workers may use temporary jobs to signal their skills by making themselves available for screening. Second, temporary jobs may provide an occasion to acquire additional human capital, social contacts and information about vacancies. It is also possible, however, that temporary employment represents a trap of endless precariousness, if it gives a 'bad signal' of lack of alternatives. There is no obvious reason to expect one or the other

* Correspondence to: Tommaso Nannicini, Department of Economics, Universidad Carlos III de Madrid, C/. Madrid 126 Office 11.2.26, 28903 Getafe, Madrid, Spain. E-mail: Tommaso.nannicini@uc3m.es
† Andrea Ichino is also affiliated with EUI, CEPR, CESifo and IZA.
Contract/grant sponsor: Italian Ministry of Welfare and the Tuscany Region.
[1] See the literature survey in OECD (2002).

mechanism to prevail. At the end of the day, whether temporary employment is a springboard or a trap is ultimately an empirical question.

With specific reference to temporary work agency (TWA) assignments, several empirical studies[2] find that these types of jobs are indeed an effective springboard toward permanent employment. All these studies share the common characteristic of using variants of the conditional independence assumption (CIA) to identify the causal effect of interest. In other words, they all use non-experimental data and assume that the selection into temporary jobs is driven by observable characteristics up to a random factor. Moreover, the vast majority of these studies make use of European data.

To the best of our knowledge, there is only one study that finds a negative effect of TWA employment on labor-market outcomes: Autor and Houseman (2005). Interestingly this study, unlike all the others, is based on a quasi-experimental setting (in the sense that it exploits a kind of randomization of treatment assignment), and, unlike the vast majority of the others, makes use of US data. Autor and Houseman argue that the evidence shown by their evaluation study is the only one we should trust, since it stems from a 'truly exogenous' source of variation in TWA assignment,[3] while the other studies, including all the available evidence about Europe, should be discarded because the CIA is likely not to hold and self-selection fully determines the positive estimates of the treatment effect.

In contrast with these statements, one may argue that both the European and the US results are valid, despite the different identification strategies, but diverge because labour markets and institutions are not the same on the two sides of the Atlantic. According to this view, it should not come as a surprise that TWA assignments have a positive effect in Europe and a negative effect in the USA on the probability of a transition to a stable job.[4] This remark does not mean that we should trust European studies only because of the existing institutional differences. It just suggests that the finding of different effects may be plausible but should be put under further scrutiny.

Using European data on TWA assignments in two Italian regions (Tuscany and Sicily), this paper proposes a sensitivity analysis for matching estimators aimed at assessing to what extent the estimates derived under the CIA are robust with respect to specific failures of this assumption. Our results show that in Tuscany a TWA assignment has a large and significant positive effect on the probability of finding a permanent job, and that this result is robust to relevant deviations from the CIA. We cannot reach the same conclusion for Sicily, where the estimated effect is positive and significant, but not robust to plausible violations of the CIA.

In light of the contraposition between European and US results mentioned above, our conclusion is that TWA jobs may have positive effects in Europe and there are institutional reasons that support this conclusion. At the same time we warn about the possibility that some of the European studies

[2] See Kvasnicka (2005), and Lechner *et al*. (2000) for Germany; Amuedo-Dorantes *et al*. (2006), and Malo and Munoz-Bullon (2002) for Spain; Anderson and Wadensjö (2004) for Sweden; Lane *et al*. (2003) for the USA; Gerfin *et al*. (2002) for Switzerland; and Booth *et al*. (2002) for the UK.

[3] More precisely, they exploit the fact that individuals applying for welfare are randomly assigned to different Work First contractors, which in turn display different policies in terms of referring their randomly assigned participants to TWAs. Hence, their identification strategy requires the additional assumption that contractors differ *only* with respect to their attitude toward TWA employment. This strategy gives rise to another peculiarity of the study by Houseman and Autor, i.e., the fact that their sample only contains low-income and at-risk workers. Moreover, what they truly estimate is the local average treatment effect (LATE), which differs from the average treatment effect on the treated (ATT) usually retrieved by other studies.

[4] Interestingly, because of the different institutional settings, also the outcome variables are slightly different in the two contexts, with the US study focusing on wages or employment duration, and the European studies focusing on the probability of attaining a permanent job.

DOI: 10.1002/jae

cited above may not be robust to violations of the CIA and thus should not be considered as evidence in favour of a springboard effect of TWA jobs.

From a methodological perspective, the sensitivity analysis for matching estimators that we propose builds on Rosenbaum and Rubin (1983a) and Rosenbaum (1987). The intuition is simple. Suppose that conditional independence is not satisfied given observables but would be satisfied if we could observe an additional binary variable. This binary variable can be simulated in the data and used as an additional matching factor in combination with the preferred matching estimator. A comparison of the estimates obtained *with* and *without* matching on this simulated variable tells us to what extent the estimator is robust to this specific source of failure of the CIA. Moreover, the simulated values of the binary variable can be constructed to capture different hypotheses regarding the nature of potential confounding factors.

Similar types of sensitivity analysis have been proposed in the literature for other kinds of estimators. For example, Rosenbaum and Rubin (1983a) and recently Imbens (2003) propose a method to assess the sensitivity of average treatment effect (ATE) estimates in parametric regression models. Altonji *et al*. (2005) use a similar idea to assess how strong selection on unobservables would have to be in order to imply that the entire estimated effect should be attributed to selection bias. However, their result is restricted to a specific parametric setup, i.e., the Heckman selection model based on the assumption of joint normality of the error terms in the selection and outcome equations. We contribute to this literature by extending this type of analysis to matching estimators of the average effect of treatment on the treated (ATT). Like Rosenbaum (1987), but differently from the above literature, we do not necessarily have to rely on any parametric model. Moreover, and unlike Rosenbaum's paper, we derive point estimates of the ATT under different possible scenarios of deviation from the CIA.

The paper is organized as follows. Section 2 presents our evaluation question within the Italian institutional context and our data collection strategy. Section 3 describes the estimation framework (i.e., propensity score matching), discusses the plausibility of its identifying assumption in our case (i.e., the CIA), and presents the baseline estimates for Tuscany and Sicily. Section 4 proposes and applies to our data a framework to assess the sensitivity of matching estimates with respect to violations of the CIA. Section 5 concludes.

2. THE CONTEXT AND THE DATA

The consequences of a TWA experience on future employment prospects have originated a very harsh debate in Italy[5] after the approval of the so-called 'Treu law' (Law 196/1997), which legalized and regulated the supply of temporary workers by authorized agencies (which were illegal until then).[6] After the introduction of this law, TWA employment has rapidly expanded,

[5] A debate which has unfortunately degenerated to the terrorist attacks that killed Massimo D'Antona in 1999 and Marco Biagi in 2002, two labour law scholars and consultants of the Ministry of Welfare. No loss of lives, fortunately, but a significant amount of social unrest has recently accompanied the proposal of introducing temporary contracts for young workers in France (the so-called CPE contract).

[6] The Treu law states that TWA employment is allowed in all but the following cases: replacement of workers on strike, firms that experienced collective dismissals in the past 12 months, and jobs that require medical vigilance. The subsequent collective agreements state that temporary workers cannot exceed 8–15% of standard employees (depending on the sector). The set of acceptable motivations includes: peak activity, one-off work, and the need for skills not available within the firm. Firms cannot sign TWA contracts with the same worker for more than four times or for a cumulated period longer than 24 months.

DOI: 10.1002/jae

especially in the north of the country and in manufacturing sectors.[7] Despite this rapid expansion and the wide interest in the debate on TWA jobs, no convincing evaluation study of their effects has yet been performed in Italy. Our paper is the first one trying to fill in this gap.[8]

In order to evaluate the effect of a single TWA assignment on the probability of finding a stable job later on, we collected data for two Italian regions, Tuscany and Sicily, which were among the few remaining areas with incomplete penetration of TWAs in 2000. In these regions, we selected five provinces that already had an agency (Livorno, Pisa and Lucca in Tuscany; Catania and Palermo in Sicily), and four that had none (Grosseto and Massa in Tuscany; Messina and Trapani in Sicily) but were otherwise similar to the previous five in terms of a wide set of economic and demographic indicators.

'Manpower Italia', a major company operating in the TWA sector, gave us the contact details of the workers in its files. From this dataset, we extracted workers who were on a TWA assignment in the nine selected provinces during the first 6 months of 2001. At that time, 'Manpower' was the only TWA company operating in these provinces, so that our dataset contains information on the universe of TWA workers in the geographic areas and in the period that we consider. This universe represents the group of treated subjects and the first 6 months of 2001 are the treatment period.[9]. Data on the treated subjects were collected through phone interviews with the computer-aided telephone interviews (CATI) method. We collected data on a random sample of control subjects drawn from the population of the nine provinces. These subjects had to satisfy two requirements: to be aged between 18 and 40, and not to have a stable job (an open-ended contract or self-employment) on 1 January 2001. This first screening of potential control subjects may be interpreted as part of the matching strategy, aimed at identifying a common support for the treated and the individuals in the comparison group, with respect to observable characteristics.

In order to reach a sufficient number of control subjects in each area, we stratified the sample according to the province of residence. Hence, our data collection strategy led to both choice-based sampling[10] and geographic stratification. It also combined *flow sampling* for the treated group and *stock sampling* for the comparison group, which may be perceived as a problem. We argue it is not for the following reasons. For TWA workers, we preferred to use flow sampling since it was the only available solution to get a sufficiently large number of treated units. For control subjects, we preferred to use stock sampling since otherwise we would have had to ask them a screening question referring to their contract in the 'prevailing part of the first 6 months of 2001', and this solution seemed a potential cause of measurement errors. Of course, our mixed sampling strategy may also create shortcomings. With respect to the alternative strategy of using flow sampling for both groups, we are incorrectly dropping from the comparison group subjects who were permanent employees on 1 January, but were temporary employees or unemployed in the first 6 months of 2001. However, it is well known that in Italy the transition probability out from standard employment is very low because of the rigidity of firing regulations. Thus, the

[7] For an aggregate picture of TWA employment in Italy, see Nannicini (2004).

[8] This evaluation study is part of a project on TWA employment financed by the Italian Ministry of Welfare and the Tuscany Region. See Ichino *et al*. (2005).

[9] Note also that the fraction of TWA workers in the total reference population is very small (around 0.6% in Tuscany and 0.2% in Sicily). This is important for the evaluation framework we adopt, because it makes the stable unit treatment value assumption (SUTVA) more plausible in our case. In fact, in this setting, it is credible to assume that treatment participation does not affect the outcome of non-participants

[10] As we will argue below (see footnote 17), this sampling strategy does not create particular problems for the matching estimation.

DOI: 10.1002/jae

group of individuals we are disregarding is likely to be very small. As a result, we believe that our sampling design is better than any feasible alternative.[11]

For both the treated and the control units, phone interviews followed an identical path of questions regarding: (a) demographic characteristics; (b) family background; (c) educational achievements; (d) work experience before the treatment period; (e) job characteristics during the treatment period; (f) work experience from the treatment period to the end of 2002; (g) job characteristics at the end of 2002. Information on the period before 1 January 2001 provided the 'pre-treatment' variables for both the treated and the control subjects, while information at the date of the interview (November 2002) provided the 'outcome' variable, defined as a binary indicator taking a value of 1 if the subject was employed with a permanent contract.

After a preliminary analysis of the data, control subjects who were out of the labour force in the treatment period (e.g., students) were dropped. Notice that this was a conservative choice with respect to the estimated treatment effects, since all these individuals had a very low probability of having a permanent job at the end of 2002. To sum up, the treated sample contains subjects who lived in the nine selected provinces and who were on a TWA assignment during the first 6 months of 2001. The comparison sample contains residents in the same provinces, aged 18–40, who belonged to the labour force but did not have a stable job on 1 January 2001 and who did not have a TWA assignment during the first 6 months of the year. The final dataset contains 2030 subjects: 511 treated and 1519 controls. Under the assumptions that will be discussed in the next section, our study aims at matching treated and control subjects in a way such that the controls can be considered as a counterfactual image of what would have happened to the treated, if they had chosen to keep looking for a stable job or to accept a different non-permanent contract in 2001.

3. THE EVALUATION FRAMEWORK

3.1. Notation

Let T be the binary variable describing treatment status: specifically, $T = 1$ if the subject was on a TWA assignment in the first 6 months of 2001, while $T = 0$ otherwise. The binary variables Y_0 and Y_1 denote the potential outcomes according to treatment status, and they take value 1 if the subject is permanently employed at the time of the interview (November 2002) and 0 otherwise. Only one of these two potential outcomes can be observed (i.e., the one corresponding to the treatment status of the subject), but the causal effect of interest is defined by their comparison: $Y_1 - Y_0$. Thus, causal inference becomes a problem of inference with missing data. In particular, we are interested in the ATT, defined as

$$E(Y_1 - Y_0|T = 1) \tag{1}$$

We assume that both the treatment status and potential outcomes are affected by a set of observable characteristics W. Our evaluation aim is to identify and consistently estimate the ATT defined in equation (1). Problems may arise because of the potential association between some of

[11] Another shortcoming of using a stock rather than a flow sample for the comparison group is that the distribution of observables is different. However, since we are only interested in the effect on the treated, this does not affect our results as matching estimators weigh comparison units so that the distribution of observables mimic that of the treated.

DOI: 10.1002/jae

the unobservable variables that affect the potential outcome in the case of no treatment and the treatment indicator T, determined itself by observable and unobservable variables. In this kind of situation, one of the assumptions that allow the identification of the ATT is 'strong ignorability' (Rosenbaum and Rubin, 1983b), which is the rationale behind common estimation strategies such as regression modelling and matching. This assumption, when the ATT is the only effect of interest, states that

$$Y_0 \perp T|W \tag{2}$$

$$Pr(T = 1|W) < 1 \tag{3}$$

Condition 2 is the already mentioned CIA, also referred to as 'unconfoundedness' or 'selection on observables' in the programme evaluation literature.[12] It means that, conditioning on observed covariates W, treatment assignment is independent of the potential outcome in the case of no treatment. The behavioural assumption behind this condition is that the untreated outcome does not influence the selection into treatment, while the possibility that self-selection depends on the treated outcome does not have to be ruled out. Although very strong, the plausibility of this assumption heavily relies on the quality and amount of information contained in W.

Condition 3 is a (weak) overlap or common-support condition. It ensures that, for each treated unit, there are control units with the same W. Under the CIA and the overlap condition, the ATT can be identified as

$$\begin{aligned} E(Y_1 - Y_0|T = 1) &= E(E(Y_1 - Y_0|T = 1, W)) \\ &= E(E(Y_1|T = 1, W) - E(Y_0|T = 0, W)|T = 1) \end{aligned} \tag{4}$$

where the outer expectation is over the distribution of W in the subpopulation of treated individuals. Thanks to the CIA, the observed outcome of control units can be used to estimate the counterfactual outcome of treated units in the case of no treatment. The next subsection provides evidence supporting the plausibility of the CIA in our specific evaluation setting.

3.2. Is the CIA Plausible in our Case?

The plausibility of the CIA crucially relies on the possibility to match treated and control units on the basis of a large and informative set of pre-treatment variables.[13] Since we were able to collect our own data, we had the opportunity to acquire information specifically designed to meet this requirement. The variables at our disposal, summarized in Table I, contain detailed information on demographic characteristics, educational attainments, family background and recent employment history of treated and control subjects. This information was collected with the same questionnaire for both the treated and controls, who were drawn from the same local labour market.[14]

Thanks to this careful data collection effort, the treated and control subjects that we consider are very similar in terms of observable characteristics at the baseline, as shown in Table I. This

[12] See Lechner (2002) and Imbens (2004).
[13] See Black and Smith (2004) for a relevant example.
[14] The importance of these two requisites for the reduction of bias when applying matching estimators is stressed by Heckman *et al*. (1997) and supported by the experimental evidence of Michalopoulos *et al*. (2004).

DOI: 10.1002/jae

Table I. Pre-treatment characteristics

	Tuscany			Sicily		
	Treated	Matched controls	All controls	Treated	Matched controls	All controls
Whole sample						
Age	26.5	27.5	29.1	26.8	27.8	30.0
Male	0.56	0.41	0.29	0.67	0.57	0.29
Single	0.90	0.87	0.66	0.83	0.81	0.49
No. of children	0.09	0.16	0.45	0.20	0.23	0.86
Father's years of schooling	9.3	9.2	8.6	8.7	9.2	7.6
Father blue-collar	0.33	0.39	0.43	0.30	0.31	0.39
Father employed	0.53	0.46	0.37	0.46	0.45	0.29
Years of schooling	12.5	12.7	12.3	12.0	12.4	11.6
Grade	75.9	77.1	76.9	74.7	74.6	76.5
Training	0.32	0.30	0.28	0.42	0.42	0.34
Distance	12.2	16.8	25.0	24.4	25.1	41.5
Unemployment period	0.38	0.42	0.48	0.42	0.44	0.62
Employed in 2000	0.35	0.36	0.42	0.34	0.35	0.30
Unemployed in 2000	0.52	0.53	0.52	0.60	0.60	0.67
Out of labour force in 2000	0.13	0.10	0.05	0.06	0.05	0.03
No. of observations	281	135	628	230	128	891
Employed in 2000						
Permanent contract	0.16	0.22	0.26	0.14	0.16	0.36
Atypical contract	0.84	0.78	0.74	0.86	0.84	0.64
Blue-collar	0.62	0.59	0.39	0.44	0.24	0.22
White-collar	0.36	0.41	0.54	0.54	0.71	0.67
Self-employed	0.02	0.00	0.07	0.01	0.04	0.10
Manufacturing	0.53	0.41	0.23	0.39	0.20	0.15
Service	0.39	0.45	0.67	0.49	0.67	0.70
Other sectors	0.08	0.14	0.11	0.11	0.13	0.15
Wage	5.2	5.6	6.8	5.6	7.6	7.0
Hours	38.0	36.3	33.3	34.5	32.1	31.1
No. of observations	98	49	266	79	45	267

Note: All the variables in this table, plus a set of dummies indicating the province of residence and birth, have been used in the estimation of the propensity score and the outcome equation. *Father blue-collar* captures whether the father has been a blue collar in the prevailing part of his working life or not. *Father employed* captures whether the father was employed in 2000 or not. *Grade* is the mark obtained in the last degree (normalized as the fraction of the highest mark for that degree). *Training* captures whether the individual received a training course in the school-to-work period or not. *Distance* is the distance from home to the nearest TWA (measured in km and calculated using ZIP codes). *Unemployment period* is the fraction of the school-to-work period spent as unemployed. *Wage* is the hourly wage in euros. *Hours* are the weekly hours of work. All the other variables—except *Age, Years of schooling*, and *Father's years of schooling*—are dummies. 'Matched controls' are individuals who belong to the control sample and are used in the nearest neighbour propensity score-matching estimation.

table reports, separately for the two regions, the average characteristics by treatment status. The differences between the two groups are arguably small, but they become even smaller when the treated subjects are compared to the matched control subjects identified with the algorithm described in Section 3.3 below. This is the subset of control subjects that are effectively used for the estimation of the causal effect of interest.

The Italian context described above provides further support for the plausibility of the CIA. Since our analysis considers provinces where TWA jobs have just appeared and we are at the very beginning of the history of TWA in Italy, we believe it is plausible to assume that, conditioning

DOI: 10.1002/jae

on our rich set of observables and particularly on the distance from the nearest agency,[15] the probability of getting in touch with an agency is the same for treated and control subjects. In other words, it is plausible that, given the recent opening of TWA in these local areas, the identity of those who enter in contact with an agency is determined by random events. As a result of this very specific situation, it becomes plausible to assume that subjects with the same observable characteristics have a different treatment status just because of chance, i.e., to assume that the CIA is satisfied.

Finally, Imbens (2004) suggests that support for the CIA can be offered by the estimation of the causal effect of a treatment that, under the CIA, is supposed not to have any effect. Not rejecting the hypothesis that a similar effect is zero would not prove that the CIA is valid, but would make this assumption considerably more plausible. We follow this suggestion by comparing, as in Heckman *et al*. (1997), two groups of control subjects that in our context can be considered respectively as 'eligible non-participants' and 'ineligible'. The first group contains subjects who declare to have contacted a temporary agency in the treatment period (first 6 months of 2001) but for whom this contact was not followed by an assignment. Thus, these subjects were eligible and potentially willing to be treated, but they were never effectively treated. The second group contains instead control subjects who had no contact with a TWA, being *de facto* equivalent to 'ineligible' individuals. Note that there is no reason to expect that the simple contact with a TWA should have any effect (under the assumption of no self-selection). Indeed, this is what we find in our data with the same methodology that we use to estimate the main causal effect of interest. Contacting a TWA without being assigned to a temporary job has an effect on the probability of finding a permanent employment equal to -0.04 (with a standard error of 0.05) in Tuscany, and equal to -0.08 (0.06) in Sicily.[16]

Needless to say, even if we find all the above arguments compelling, we are aware of the possibility that the CIA might fail in several ways in our context. Precisely for this reason, in Section 4, we propose a sensitivity analysis in order to assess the robustness of our estimates to specific violations of the CIA. Before doing so, however, we present our matching strategy and the baseline results in the next two subsections.

3.3. Propensity Score Matching

Since many of the covariates W summarized in Table I are multivalued or continuous, some smoothing techniques are in order. Under the CIA, several estimation strategies can serve this purpose. One of these is regression modelling. Using regression to 'adjust' or 'control for' pre-intervention covariates is, in principle, a good strategy, although it has some pitfalls. For instance, if there are many covariates, as in our case, it can be difficult to find an appropriate specification. Moreover, regression modelling obscures information on the distribution of covariates in the two treatment groups. In principle, one would like to compare individuals that have the same values of all covariates. Unless there is a substantial overlap of the two distributions of covariates, with

[15] The distance measure affects both the treatment assignment and the outcome, under the credible assumption that, within each province, TWAs locate in the area with higher labour demand. It is thus important to control for this variable in order to capture local-market effects that are observable to the TWA but not to the econometrician.

[16] Even if these point estimates are insignificant, they both have negative sign, suggesting the possibility of negative, if any, self-selection. This would even make the ATT estimates presented in Section 3.4 a lower bound of the true causal effect of a TWA assignment.

DOI: 10.1002/jae

regression one has to rely heavily on model specification (i.e., on extrapolation) for the estimation of treatment effects. It is thus crucial to check how much the two distributions overlap and what is their 'region of common support'. When the number of covariates is large, this task is not an easy one. A possible solution is to reduce the problem to a single dimension by using propensity score-matching techniques.

The propensity score is the individual probability of receiving the treatment given the observed covariates: $p(W) = P(T = 1|W)$. Under the CIA, Y_0 and Y_1 are independent of T given $p(W)$ (Rosenbaum and Rubin, 1983b). Note that the propensity score satisfies the so-called 'balancing property', i.e., observations with the same value of the score have the same distribution of observable characteristics irrespective of treatment status; moreover, the exposure to treatment or control status is random for a given value of the score. These properties allow the use of the propensity score as a univariate summary of all W.

If $p(W)$ is known, the ATT can be estimated as follows:

$$\begin{aligned}\tau &\equiv E(Y_1 - Y_0|T = 1) = E(E(Y_1 - Y_0|p(W), T = 1)) \\ &= E(E(Y_1|p(W), T = 1) - E(Y_0|p(W), T = 0)|T = 1)\end{aligned} \tag{5}$$

where the outer expectation is over the distribution of $(p(W)|T = 1)$. Any probabilistic model can be used to estimate the propensity score, as long as the resulting estimate satisfies the properties that the propensity score should have. We assume $Pr(T = 1|W) = F(h(W))$, where $F(.)$ is the normal cumulative distribution and $h(W)$ is a function of the covariates with linear and higher-order terms.[17] Since the specification of $h(W)$ which satisfies the balancing property is more parsimonious than the full set of interactions needed to match treated and control units according to observables, the propensity score reduces the dimensionality problem of our matching strategy.

We estimate the propensity score separately in Tuscany and Sicily. The propensity score specification for each region includes all the pre-treatment variables mentioned in Table I.[18] Even though the treated and the comparison groups are spread around the whole region of the common support, for high values of the propensity score the relative size of the controls is very small if compared with the treated.[19] This means that we end up using these very few control subjects to estimate the 'counterfactual' outcome of the treated in this block, with the risk of obtaining sensitive results. Hence, to assess the robustness of the estimates with respect to the intensity with which the upper tail of the comparison group gets used, we also estimate the ATT in a region of 'thick support', as proposed by Black and Smith (2004).

The final step of our estimation strategy is the use of the nearest neighbour algorithm to identify the best match for each treated subject, given that the probability of observing two

[17] In a choice-based sampling scheme, which is the scheme of many empirical studies like the one presented in this paper, the odds ratio of the misspecified (i.e., choice-based) propensity score can be used to implement matching as suggested by Heckman and Todd (1999). The misspecified odds ratio is monotonically related to the odds ratio of the true propensity score, which is itself a monotonic transformation of the score. Note, however, that the CIA holds also in the choice-based sample, although the true propensity score cannot be consistently estimated. Hence, as long as the balancing property is satisfied, the choice-based score can still be used as a balancing score, in order to construct a comparison group with the same distribution of covariates as the treated group.

[18] In Tuscany, to have the balancing test satisfied we also included the interaction term between self-employment and one of the provinces, and the squared distance.

[19] For values of $p(W)$ higher than 0.80, in Tuscany (Sicily) there are only 6 (7) comparison units against 43 (17) treated units. See the previous working paper version (Ichino *et al*., 2006) for further details on the propensity score estimation.

DOI: 10.1002/jae

units with exactly the same value of the (continuous) score is in principle zero.[20] The nearest neighbour algorithm compares each treated unit with the comparison unit that is closest in terms of the propensity score. Because we allow for replacement, a single control can be the best match of more than one treated unit. Since Abadie and Imbens (2006) show that the bootstrap variance estimator is invalid for nearest neighbour matching, we calculate analytical standard errors assuming independent outcomes across units.

3.4. The effect of TWA Assignments in Tuscany and Sicily

Table II presents the estimation results obtained with the matching strategy described above. The first row contains the baseline matching estimates for the whole sample. The ATT is estimated to be equal to 19 percentage points in Tuscany, with a standard error of 0.06. In this region, the observed probability to have a stable job in the outcome period is 31% for the treated and 17% for the controls (see Table I). Thus, our matching strategy increases by 5 percentage points the estimated effect of a TWA job with respect to what would be implied by the naive comparison of the raw statistics of treated and controls.[21] Note also that the estimated 'counterfactual' probability to get a permanent job for the treated in case of no treatment is 12% (i.e., 31 minus 19). This estimated probability is 5 percentage points lower than the average probability observed for all control subjects. This indicates that the treated tend to be subjects who would have worse-than-average employment opportunities in the absence of a TWA assignment. These are the workers for whom TWA jobs may be an attractive option. In Sicily, the baseline ATT estimate is equal to 10 percentage points, with a standard error of 0.05. In this region, the springboard effect of a TWA assignment is weaker but still significant.[22]

The second row of Table II presents ATT estimates obtained considering only subjects with a propensity score in the region of 'thick support', i.e., for $p(W) \in (0.33, 0.67)$. This is suggested by Black and Smith (2004) to assess the robustness of estimates with respect to the frequency of control subjects in the upper tail of the comparison group. Both in Tuscany and in Sicily, the thick-support estimates are greater than the baseline ones. Since self-selection is likely to hit the region of 'thin support' more than the region of thick support, the fact that the point estimates in the latter region do not fall is another supporting element for the claim that self-selection is not driving the baseline results.

Table II also presents some heterogeneity results. In Tuscany, the estimated ATT is greater for males and for individuals older than 30, even though for females and individuals under 30 the effect is never lower than 10 percentage points. In Sicily, the estimated ATTs for males and individuals younger than 30 are similar to the baseline estimate of 0.10, while the ATTs for females and individuals older than 30 are completely insignificant.

To sum up, the nearest neighbour propensity score-matching estimates, based on the CIA, detect a positive and significant 'springboard' effect of TWA employment in the Italian context. This effect is larger in one region (Tuscany, 0.19) than in the other (Sicily, 0.10), but overall this

[20] We only present results based on the nearest neighbour algorithm, which is simple and intuitive, given that all the other algorithms that we implemented (specifically, kernel and stratification) produced similar ATT estimates. See Dehejia and Wahba (2002), Smith and Todd (2005), or Caliendo and Kopeinig (2007) for a presentation of different matching algorithms.

[21] The OLS estimate obtained conditioning on the same set of covariates and using all the control subjects is equal to 0.16 with a standard error of 0.03.

[22] In this case, the OLS estimate is equal to 0.05 (0.03).

DOI: 10.1002/jae

Table II. Effect of a TWA assignment on the probability to find a permanent job: Nearest neighbour propensity score matching

	Tuscany			Sicily		
	ATT	Treated	Controls	ATT	Treated	Controls
Whole sample	0.19	281	133	0.10	230	131
	(0.06)			(0.05)		
Thick-support	0.23	109	56	0.14	92	43
	(0.07)			(0.08)		
Male	0.24	157	59	0.10	155	76
	(0.10)			(0.07)		
Female	0.14	124	71	−0.07	75	57
	(0.07)			(0.06)		
Under 30	0.11	199	88	0.09	170	90
	(0.07)			(0.06)		
Over 30	0.33	82	44	0.00	60	39
	(0.09)			(0.09)		

Note: Standard errors in parentheses. The first-row ATT is the baseline estimate for the whole sample. The 'thick-support' estimation considers only the observations with an estimated propensity score in the region (0.33, 0.67). The ATTs for males, females, and individuals under 30 and over 30 are estimated separately in these subsamples. The number of controls refers to the matched controls used by the nearest neighbour algorithm.

evidence concerning Italy sustains the main findings of the European studies on TWA jobs, i.e., that this kind of non-standard employment relationship is able to improve the future labour market outcomes of workers. Nevertheless, even if the arguments proposed in Section 3.2 to support the validity of the CIA appear convincing, we believe that a sensitivity analysis like the one described in the next section is needed to decide confidently whether these estimates can be trusted or not.

4. SENSITIVITY ANALYSIS

In this section, we describe a sensitivity analysis aimed at assessing the bias of ATT estimates when the CIA is assumed to fail in some specific and meaningful ways. We suggest that this kind of sensitivity analysis should always accompany the presentation of matching estimates obtained under the CIA. Note, however, that what we propose is not a 'test' of the CIA. Indeed, this identifying assumption is intrinsically non-testable because the data are uninformative about the distribution of Y_0 for treated units. Nevertheless, the sensitivity analysis that we propose provides valuable information in order to draw conclusions on the reliability of matching estimates.

4.1. Our Proposal and the Related Literature

We build on the work of Rosenbaum and Rubin (1983a), who propose assessing the robustness of the estimated causal effects (in particular, the ATE) with respect to assumptions about an unobserved binary covariate that is associated with both the treatment and the response. The unobservables are assumed to be summarized by a binary variable in order to simplify the analysis, although similar techniques could be used assuming some other distribution for the unobservables. The central assumption of their analysis is that the assignment to treatment is not unconfounded

DOI: 10.1002/jae

given the set of observable variables W, i.e.,

$$Pr(T = 1|Y_0, Y_1, W) \neq Pr(T = 1|W) \tag{6}$$

but the CIA holds given W and an unobserved binary covariate U:

$$Pr(T = 1|Y_0, Y_1, W, U) = Pr(T = 1|W, U) \tag{7}$$

Given these assumptions, which are common to all of the other sensitivity analysis methods discussed below, Rosenbaum and Rubin (1983a) suggest specifying four (sets of) parameters that characterize the distribution of U and the association of U with T, Y_1 and Y_0 given observed covariates (or given strata defined by observed covariates). The unobservable U is usually assumed to be independent of the observed covariates, i.e., $Pr(U = 1|W) = Pr(U = 1)$. After this step, the full likelihood for T, $Y_0, Y_1, U|W$ is derived and maximized, holding the sensitivity parameters as fixed known values. It is then possible to judge the sensitivity of inferential conclusions with respect to certain plausible variations in the assumptions about the association of U with T, Y_0 and Y_1. If conclusions are relatively insensitive over a range of plausible assumptions about U, causal inference is more defensible.

Imbens (2003) applies the same method but expresses the sensitivity parameters in terms of partial R^2, in order to ease the interpretation of results. Note, however, that the approach followed by these authors uses a parametric model as the basis for the estimation of the average treatment effects: specifically, a normal model when the outcome is continuous as in Imbens (2003) and a logistic regression when the outcome is binary as in Rosenbaum and Rubin (1983a). Parameterization is instead not necessary in the sensitivity analysis that we propose in this paper.

Rosenbaum (1987) proposes assessing the sensitivity of significance levels and confidence intervals, rather than the sensitivity of point estimates. The method involves only one sensitivity parameter (which represents the association of T and U), instead of the four (sets of) sensitivity parameters specified in Rosenbaum and Rubin (1983a), so that the joint distributions of $T, Y_1, U|W$ and $T, Y_0, U|W$ are only partially specified. As a consequence, only bounds for significance levels and confidence intervals can be derived.

Our proposed method aims instead at assessing the sensitivity of point estimates (and specifically the sensitivity of ATT matching estimates). Like Rosenbaum (1987), we do not rely on any parametric model for the outcome, but, unlike his paper, we derive point estimates of the ATT under different possible scenarios of deviation from the CIA. Instead of estimating by maximum likelihood a model for the outcome and the treatment status involving the confounding factor U, we impose the values of the parameters that characterize the distribution of U. Given these parameters, we then predict a value of the confounding factor for each treated and control subject and we re-estimate the ATT including the simulated U in the set of matching variables. By changing the assumptions about the distribution of U, we can assess the robustness of the ATT with respect to different hypotheses regarding the nature of the confounding factor. Moreover, we can verify whether there exists a set of plausible assumptions on U under which the estimated ATT is driven to zero by the inclusion of U in the matching set.

More formally, we consider for expositional simplicity the case of binary potential outcomes $Y_0, Y_1 \in \{0, 1\}$, as in the analysis of the effect of TWAs in Italy discussed in Section 3, and we denote by $Y = T \cdot Y_1 + (1 - T) \cdot Y_0$ the observed outcome for a given unit, which is equal to one

DOI: 10.1002/jae

of the two potential outcomes depending on treatment exposure.[23] Assuming that equations (6) and (7) are satisfied (with the latter representing the extended CIA in the new setting), we characterize the distribution of the unobserved binary confounding factor U by specifying the parameters

$$Pr(U = 1|T = i, Y = j, W) = Pr(U = 1|T = i, Y = j) \equiv p_{ij} \tag{8}$$

with $i, j \in \{0, 1\}$, which give the probability that $U = 1$ in each of the four groups defined by the treatment status and the outcome value.[24]

Given arbitrary (but meaningful) values of the parameters p_{ij}, our sensitivity analysis proceeds by attributing a value of U to each subject, according to her belonging to one of the four groups defined by the treatment status and the outcome. We then treat U as any other observed covariate and, in particular, we include U in the set of matching variables used to estimate the propensity score and to compute the ATT according to the nearest neighbour estimator. Using a given set of values of the sensitivity parameters, we repeat the matching estimation many times (i.e., $m = 1000$) and obtain an estimate of the ATT, which is an average of the ATTs over the distribution of the simulated U. Thus, for any given configuration of the parameters p_{ij}, we can retrieve a point estimate of the ATT which is robust to the specific failure of the CIA implied by that configuration.

Despite its simplicity, this sensitivity analysis has several advantages. First, note that the hypothesized associations of U with Y and T are stated in terms of proportions characterizing the distribution of $U|T, Y, W$. This avoids a possibly incorrect parametric specification of the distribution of $Y|T, U, W$, which is the strategy adopted by competing types of sensitivity analysis. For example, Altonji *et al.* (2005) use a standard selection model as a benchmark for their sensitivity analysis. In this way they can obtain some analytical results at the cost of imposing a model that assumes a constant (in the logit scale) treatment effect and a single parameter (the correlation between the error terms in the selection and outcome equations) to characterize both the unobserved selection into treatment and its association with the outcome.

Second, the parameters p_{ij} (which in turn determine the parameters $p_{i\cdot}$) can be chosen to make the distribution of U similar to the empirical distribution of observable binary covariates. In this case, the simulation exercise reveals the extent to which matching estimates are robust to deviations from the CIA induced by the impossibility of observing factors similar to the ones used to calibrate the distribution of U. This is a different exercise from the simple removal of an observed variable from the matching set W, since in our simulations we are still controlling for all the relevant covariates observed by the econometrician. Third, one can search for the existence of a set of parameters p_{ij} and $p_{i\cdot}$ such that if U were observed the estimated ATT would be driven to zero, and then assess the plausibility of this configuration of parameters. If all of the configurations leading to such a result could be considered very unlikely, the exercise would support the validity

[23] Note that this sensitivity analysis can be adapted to multi-valued or continuous outcomes by simulating U on the basis of a binary transformation of the outcome. See Nannicini (2007) for an example.

[24] Note that using these parameters and the probability of a given outcome by treatment status $Pr(Y = i|T = j)$, which is observed in the data, we can compute the fraction of subjects with $U = 1$ by treatment status only:

$$p_{i\cdot} \equiv Pr(U = 1|T = i) = \sum_{j=0}^{1} p_{ij} \cdot Pr(Y = j|T = i) \quad i \in \{0, 1\}$$

Hence, by setting the parameters p_{ij}, we can generate situations in which the fraction of subjects with $U = 1$ is greater among the treated ($p_{1\cdot} > p_{0\cdot}$) or among controls ($p_{1\cdot} < p_{0\cdot}$).

DOI: 10.1002/jae

of the estimates derived under the CIA. Finally, our simulation-based sensitivity analysis is capable of assessing the robustness of matching estimates of the ATT irrespective of the specific algorithm used to match observations.

4.2. Interpretation of the Sensitivity Parameters

Equation (8) assumes that the distribution of U given T and Y does not vary with W. In principle one could relax this assumption if there existed an obvious way to model explicitly the association between U and other important covariates like gender or education. Even if such a route were feasible in our case, however, taking it would not be strictly necessary because this simplifying assumption concerning the irrelevance of W in the simulation of U does not alter the interpretation of the sensitivity parameters.

To understand the problem, note that the threat to the baseline ATT estimate comes from the possibility that $Pr(Y_0 = 1|T, W, U) \neq Pr(Y_0 = 1|T, W)$, which implies that, without observing U, the outcome of control subjects cannot be used to estimate consistently the counterfactual potential outcome of the treated in case of no treatment. As a result, it would seem that the parameters p_{ij}, which fully determine the distribution of the simulated U, cannot be used to simulate a similar confounder because they are defined disregarding W and refer to the observed outcome of control subjects, not to the potential untreated outcome. However, it can be shown that[25]

$$\begin{aligned} p_{01} > p_{00} &\Rightarrow Pr(U = 1|T = 0, Y = 1, W) > Pr(U = 1|T = 0, Y = 0, W) \\ &\Rightarrow Pr(Y = 1|T = 0, U = 1, W) > Pr(Y = 1|T = 0, U = 0, W) \end{aligned}$$

Moreover, under the extended CIA (i.e., under the assumption that assignment to treatment is unconfounded given both W and U), we also have that

$$\Rightarrow Pr(Y_0 = 1|T = 0, U = 1, W) > Pr(Y_0 = 1|T = 0, U = 0, W)$$

Hence, by simply assuming that $p_{01} > p_{00}$, we can simulate a confounding factor that has a positive effect on the potential outcome in case of no treatment, disregarding how this confounding factor might be correlated with W. The same chain of inequalities (and reasoning) applies to the assumption that $p_{1\cdot} > p_{0\cdot}$, which can be imposed by setting p_{11} and p_{10} appropriately.

These results allow us to interpret the sensitivity parameters p_{ij} and $p_{i\cdot}$ in a meaningful way even without modelling explicitly the relationship between U and W, and even if we focus on the observed outcome of control subjects and not on their potential outcome. This is because the real threat to the baseline estimate is coming from a potential confounder that has both a positive effect on the untreated outcome ($p_{01} - p_{00} > 0$) and on the selection into treatment ($p_{1\cdot} - p_{0\cdot} > 0$).[26] The presence of such a confounder, even without a true causal relationship between T and Y, could completely determine a positive ATT estimate. As a consequence, the sensitivity simulations should focus precisely on confounders of this type.

If the above analysis solves the problem of simulating the *sign* of the effects of a potential confounder, it is not enough to solve the problem of measuring the *size* of these effects. As a

[25] See the previous working paper version (Ichino *et al*., 2006) for a formal proof.

[26] This kind of reasoning assumes a positive baseline estimate, but since the treatment is binary this is just a matter of definition.

DOI: 10.1002/jae

matter of fact, one might be tempted to interpret the difference $d = p_{01} - p_{00}$ as a measure of the effect of U on the untreated outcome, and the difference $s = p_{1\cdot} - p_{0\cdot}$ as a measure of the effect of U on treatment assignment. But these effects must be evaluated after conditioning on W because even if the distribution of U given T and Y does not vary with W, there will be in the data an association between U and W, coming indirectly from the association of W with T and Y.

To sidestep this shortcoming, we implement the sensitivity analysis by measuring how the different configurations of p_{ij} chosen to simulate U translate into associations of U with Y_0 and T (conditioning on W). More precisely, by estimating a logit model of $Pr(Y = 1|T = 0, U, W)$ in every iteration, we can compute the effect of U on the relative probability to have a positive outcome in case of no treatment (the observed 'outcome effect' of the simulated U) as the average estimated odds ratio of the variable U:

$$\frac{\dfrac{P(Y=1|T=0,U=1,W)}{P(Y=0|T=0,U=1,W)}}{\dfrac{P(Y=1|T=0,U=0,W)}{P(Y=0|T=0,U=0,W)}} \equiv \Gamma$$

Similarly, by estimating the logit model of $Pr(T = 1|U, W)$, the average odds ratio of U would measure the effect of U on the relative probability to be assigned to the treatment $T = 1$ (the observed 'selection effect' of U):

$$\frac{\dfrac{P(T=1|U=1,W)}{P(T=0|U=1,W)}}{\dfrac{P(T=1|U=0,W)}{P(T=0|U=0,W)}} \equiv \Lambda$$

By simulating U under the assumptions that $d > 0$ and $s > 0$, both the outcome and the selection effect must be positive (i.e., Γ and Λ must be greater than one). Moreover, from a quantitative point of view there should be a non-monotonic but close relationship between d and Γ and between s and Λ. Hence, by simulating U on the basis of the parameters p_{ij} and by displaying the associated Γ and Λ, we can perform an informative sensitivity analysis even without modelling the association between U and W.

Finally, note that, in order to assess the relevance of the above simulation assumptions (in particular, the fact that U is binary and that it does not depend on W), in a previous working paper version (Ichino *et al*., 2006) we also performed Monte Carlo exercises aimed at evaluating how the sensitivity analysis that we propose would change in response to different data-generating processes (DGPs). A first set of Monte Carlo exercises showed that the assumption of a binary confounder, when the true one is continuous, tends to produce conservative ATT estimates. As a result, with respect to this modelling assumption concerning U, the sensitivity analysis that we propose should not lead to infer that the ATT estimates are robust to failures of the CIA when in fact they are not. This result is consistent with the finding by Wang and Krieger (2005) that causal conclusions are more sensitive to unobserved binary covariates than (normal) continuous unobservables. In a second set of Monte Carlo exercises, we simulated data in which the CIA holds given W and a binary U, but assumed that U depends on some relevant W in the 'true' DGP and not in the sensitivity analysis. The results showed the robustness of the analysis also with respect to the simulation assumption that $U|Y, T$ does not depend on W.

DOI: 10.1002/jae

4.3. Sensitivity and Bounds

The sensitivity analysis that we propose starts from a point-identifying assumption (the CIA in our case) and then examines how the results change as this assumption is weakened in specific ways. A complementary approach, proposed by Manski (1990), consists of dropping the CIA entirely, and constructing bounds for the treatment effect that rely on either the outcome being bounded or on alternative identifying assumptions. It is useful to clarify with an example the relationship between these two approaches.

Consider the ATT defined as follows:

$$\text{ATT} = E(Y_1|T = 1) - E(Y_0|T = 1) \tag{9}$$

As already noted, because Y_0 is not observed when $T = 1$, the term $E(Y_0|T = 1)$ cannot be estimated from the data alone. Nevertheless if Y_0 is, for example, a binary variable taking the value 1 or 0 (or more generally a bounded variable), one can obtain non-parametric bounds for the ATT, substituting $E(Y_0|T = 1)$ with its smallest and largest possible values:

$$E(Y_1|T = 1) - 1 \leq ATT \leq E(Y_1|T = 1) \tag{10}$$

These bounds can be estimated using sample analogues. Our sensitivity analysis offers a way to understand what set of assumptions concerning a potential confounder U would lead to an ATT equal to the lower or the upper non-parametric bound.

It is easy to show that the lower bound is achieved when, among the treated, there are only individuals with $U = 1$, i.e., $Pr(U = 1|T = 1) = 1$, and among the controls all the individuals with $U = 1$ have $Y_0 = 1$, i.e., $Pr(Y_0 = 1|T = 0, U = 1) = 1$. This translates into the following assumptions on the parameters p_{ij} : $p_{11} = 1$; $p_{10} = 1$; $p_{01} = k > 0$; $p_{00} = 0$.

The upper bound is instead achieved when, among the treated, there are only individuals with $U = 1$, i.e., $Pr(U = 1|T = 1) = 1$, and among the controls all the individuals with $U = 1$ have $Y_0 = 0$, i.e., $Pr(Y_0 = 1|T = 0, U = 1) = 0$. This translates into the following assumptions on the parameters p_{ij} : $p_{11} = 1$; $p_{10} = 1$; $p_{01} = 0$; $p_{00} = k > 0$.

These sets of circumstances are really extreme and thus seem highly implausible. This explains why non-parametric bounds are often uninformative in specific applications. This happens because there exist sets of values of the treatment effect which are within the bounds but correspond to scenarios that are very unlikely, despite being potentially possible. Thanks to reasonable assumptions on the association between confounding factors, treatment status and potential outcomes, a sensitivity analysis like the one proposed in this paper offers the possibility of restricting the size of the non-parametric bounds by eliminating possible but unlikely values of the ATT.

4.4. Results of the Sensitivity Analysis

We are now ready to show how the sensitivity analysis proposed above can complement in a useful way the empirical results presented in Section 3. Tables III and IV display the basic results for Tuscany. For expositional simplicity, let us say that U measures some unobservable component of ability, which for brevity we call 'skill'. Each row of the first four columns of Table III contains the four probabilities $p_{ij} = Pr(U = 1|T = i, Y = j)$, with $i, j \in \{0, 1\}$, which characterize the binary

DOI: 10.1002/jae

distribution of skill, by treatment status and outcome, under which the ATT has been estimated. Hence, for example, p_{11} indicates the fraction of skilled subjects among those who are treated and find a permanent job after treatment, and so on. The estimated Γ provides an indication of the 'outcome effect' of U, i.e., the effect of skill on the untreated outcome, controlling for the observable covariates W. Similarly, the estimated Λ measures the 'selection effect' of U, i.e., its effect on the assignment to treatment, again controlling for observables.

To facilitate a comparison between actual and simulated results, the first row of Table III shows the baseline ATT estimate obtained with no confounder in the matching set. The second row reports the ATT estimated with a neutral confounder (i.e., one such that $d = p_{01} - p_{00} = 0$ and $s = p_{1.} - p_{0.} = 0$): such a confounder is enough to slightly perturbate the baseline result.[27] The other rows of Table III show how the baseline estimate changes when the binary confounding factor U is calibrated to mimic different observable covariates and is then included in the set of matching variables.

Table III. Sensitivity analysis in Tuscany: effect of 'calibrated' confounders

	Fraction $U = 1$ by treatment/outcome				Outcome effect Γ	Selection effect Λ	ATT	SE
	p_{11}	p_{10}	p_{01}	p_{00}				
No confounder	0.00	0.00	0.00	0.00	—	—	0.19	0.06
Neutral confounder	0.50	0.50	0.50	0.50	1.00	1.00	0.16	0.07
Confounder-like								
Male	0.55	0.56	0.32	0.28	1.2	3.3	0.15	0.07
Single	0.86	0.92	0.76	0.64	2.0	5.1	0.15	0.07
High school	0.75	0.74	0.69	0.71	0.9	1.2	0.16	0.07
University	0.14	0.12	0.13	0.19	0.6	0.7	0.15	0.07
Prev. employed	0.40	0.33	0.50	0.41	1.5	0.7	0.16	0.06
Prev. permanent	0.08	0.05	0.25	0.08	4.5	0.5	0.16	0.07
Manufacturing	0.19	0.18	0.23	0.09	1.7	2.3	0.14	0.07
Father educ.	0.34	0.31	0.32	0.27	1.3	1.3	0.16	0.07
High distance	0.20	0.14	0.54	0.49	1.3	0.2	0.17	0.07

Note: Let U be a binary confounding factor and denote the fraction of $U = 1$ by treatment and outcome as: $p_{ij} = Pr(U = 1|T = i, Y = j)$, with $i, j = \{0, 1\}$. On the basis of these parameters, a value of U is imputed to each individual and the ATT is estimated by nearest neighbour propensity score matching with U in the set of matching variables. The process is repeated 1000 times. Γ is the average estimated odds ratio of U in the logit model of $Pr(Y = 1|T = 0, U, W)$; Λ is the average estimated odds ratio of U in the logit model of $Pr(T = 1|U, W)$; 'ATT' is the average of the simulated ATTs; 'SE' is the standard error (calculated as shown in equation (11)). The first two rows show the ATT estimate with no confounding factor or with a confounder whose outcome and selection effects are insignificant, respectively. In the 'confounder-like' rows, U has been calibrated to match the distribution of the corresponding covariate.

[27] In order to compute a standard error of the ATT estimator when U is included in the set of matching variables, we considered the problem of the unobserved confounding factor as a problem of missing data that can be solved by multiply imputing the missing values of U (Rubin, 1987). Let m be the number of imputations (i.e., replications) of the missing Us, and let $\hat{ATT}_k$ and se_k^2 be the point estimate and the estimated variance of the ATT estimator at the kth imputed data set, $k = 1, 2, \ldots, m$. The ATT estimate, $\hat{ATT}$, is then obtained (as already explained in Section 4.1) by the average of the $\hat{ATT}_k$s over the m imputations. As we showed in the previous working paper version (Ichino *et al.*, 2006), the total variance associated with $\hat{ATT}$ can be estimated as

$$T = \frac{1}{m}\sum_{k=1}^{m} se_k^2 + \frac{m+1}{m(m-1)}\sum_{k=1}^{m}(\hat{ATT}_k - \hat{ATT})^2 \tag{11}$$

DOI: 10.1002/jae

Table IV. Sensitivity analysis in Tuscany: characterizing 'killer' confounders

	$s = 0.1$ $\Lambda \in$ [1.5, 1.6]	$s = 0.2$ $\Lambda \in$ [2.4, 2.5]	$s = 0.3$ $\Lambda \in$ [3.8, 4]	$s = 0.4$ $\Lambda \in$ [6.1, 6.4]	$s = 0.5$ $\Lambda \in$ [9.9, 10.3]	$s = 0.6$ $\Lambda \in$ [18.9, 20]	$s = 0.7$ $\Lambda \in$ [42, 45.4]
$d = 0.1$ $\Gamma \in$ [1.6, 1.9]	0.16	0.15	0.14	0.14	0.13	0.12	0.11
	(0.07)	(0.07)	(0.08)	(0.09)	(0.09)	(0.11)	(0.14)
$d = 0.2$ $\Gamma \in$ [2.5, 3.5]	0.15	0.14	0.12	0.11	0.10	0.08	0.04
	(0.07)	(0.08)	(0.08)	(0.09)	(0.11)	(0.12)	(0.16)
$d = 0.3$ $\Gamma \in$ [3.9, 6]	0.14	0.12	0.10	0.09	0.06	0.03	−0.02
	(0.07)	(0.08)	(0.08)	(0.09)	(0.11)	(0.14)	(0.17)
$d = 0.4$ $\Gamma \in$ [6.5, 9.7]	0.14	0.11	0.09	0.06	0.03	−0.01	−0.06
	(0.07)	(0.08)	(0.09)	(0.10)	(0.12)	(0.14)	(0.18)
$d = 0.5$ $\Gamma \in$ [11.8, 18.2]	0.13	0.09	0.07	0.04	0.00	−0.06	−0.12
	(0.07)	(0.08)	(0.09)	(0.10)	(0.12)	(0.15)	(0.19)
$d = 0.6$ $\Gamma \in$ [23, 36.7]	0.12	0.08	0.05	0.01	−0.03	−0.10	−0.19
	(0.07)	(0.09)	(0.09)	(0.11)	(0.13)	(0.16)	(0.20)
$d = 0.7$ $\Gamma \in$ [55.1, 81]	0.12	0.07	0.02	−0.02	−0.07	−0.13	−0.23
	(0.07)	(0.09)	(0.10)	(0.12)	(0.14)	(0.17)	(0.21)

Note: Under the assumption that $Pr(U = 1) = 0.4$ and $p_{11} - p_{10} = 0$, the differences $d = p_{01} - p_{00}$ (which captures the outcome effect of U in the absence of treatment) and $s = p_{1\cdot} - p_{0\cdot}$ (which captures the effect of U on the selection into treatment) uniquely define the parameters p_{ij}, with $i, j = \{0, 1\}$. In each cell, the simulated ATT associated to the corresponding differences is reported (standard errors in parentheses). All ATTs are averaged over 1000 iterations. Γ is the average estimated odds ratio of U in the logit model of $Pr(Y = 1|T = 0, U, W)$; Λ is the average estimated odds ratio of U in the logit model of $Pr(T = 1|U, W)$. The baseline estimate without confounder is equal to 0.19 in Tuscany. The Manski bounds are $(-0.69, 0.31)$.

The first case sets the distribution of U to be similar to the distribution of gender. In this case, given that 55% of the subjects who are exposed to treatment and find a permanent job are male, by setting $p_{11} = 0.55$ we impose that an identical fraction of subjects are skilled and therefore are assigned a value of U equal to 1. An analogous interpretation holds for the other probabilities p_{ij} in this row. Note that these assumptions imply that the treated are more skilled than the controls in the whole sample. When controlling for observables, skill has a slightly positive effect on the relative probability of getting a permanent job in case of no treatment ($\Gamma = 1.2 > 1$) and a much higher effect on the relative probability of being treated ($\Lambda = 3.3 > 1$). Under a deviation from the CIA with these characteristics, the ATT is estimated to equal 0.15. This estimate differs by only four (one) percentage points with respect to the baseline (neutral) estimate obtained in the absence of confounding effects, and remains statistically significant.

The other rows assume that the distribution of U is in turn comparable to the distribution of observables like marital status, high school degree, university degree, existence of previous work experience or of a previous permanent contract, previous job in manufacturing, high education of the father, and living far away from a TWA. All these variables have a significant role either in the propensity score estimation or in the outcome equation. Only in the case of skill behaving like a previous job in manufacturing (associated with an outcome effect of $\Gamma = 1.7$ and a selection effect of $\Lambda = 2.3$) does the ATT differ by five (two) percentage points from the baseline (neutral) estimate, but it still remains statistically (and economically) significant.

Taken in conjunction, these simulations convey an impression of robustness of the baseline matching estimate of the ATT in Tuscany. These simulations also show that both the outcome and the selection effect of U must be strong in order to represent a threat to the significance of the estimated ATT. The advantage of our sensitivity exercise, however, goes beyond these findings,

DOI: 10.1002/jae

because it allows us to explore the characteristics of the confounding factor U under which the point estimate of the ATT becomes close to zero. This is done in Table IV. To reduce the dimensionality of the problem in the search for a characterization of 'killer' confounding factors, we fix at some predetermined values the parameters $Pr(U = 1)$ and $p_{11} - p_{10}$. The former represents the fraction of skilled individuals in the whole sample, while the latter captures the effect of skill on the treated outcome. Since these parameters are not expected to represent a threat for the estimated ATT, we can keep them at fixed known values and fully characterize the simulated confounder by varying the already defined differences: $d = p_{01} - p_{00}$ and $s = p_{1.} - p_{0.}$. We can follow this route because the difference $(p_{11} - p_{10})$ is fixed, and $Pr(U = 1)$ can be expressed as

$$\begin{aligned} Pr(U = 1) = {} & p_{11} \cdot Pr(Y = 1|T = 1) \cdot Pr(T = 1) + p_{10} \cdot Pr(Y = 0|T = 1) \cdot Pr(T = 1) \\ & + p_{01} \cdot Pr(Y = 1|T = 0) \cdot Pr(T = 0) + p_{00} \cdot Pr(Y = 0|T = 0) \cdot Pr(T = 0) \end{aligned}$$

As a result, we have a system of four equations[28] that allows us to retrieve the four parameters p_{ij} and simulate the confounding factor uniquely associated with the preferred values of d and s.

A further problem results from the fact that, as discussed in Section 4.2, the differences d and s are set without taking into account the role of W. However, we have shown that we can associate the values of d and s with the parameters Γ and Λ, respectively. These estimated odds ratios provide a measure of the observed effect of the confounder U on the outcome and on the selection into treatment (controlling for W). Table IV shows the results of this simulation exercise for Tuscany. The fraction of skilled individuals in the whole sample $Pr(U = 1)$ is assumed to be equal to 0.04, while the effect of skill on the treated outcome $(p_{11} - p_{10})$ is normalized to zero.[29]

Along every row of Table IV, d is kept fixed while s is increasing. Along every column, the opposite happens. In each row, the predetermined value of d is associated with the range of variation of the estimated outcome effect Γ that characterizes the corresponding simulated confounders. Similarly, in each column, the value of s is associated with the range of variation of the estimated selection effect Λ that characterizes the corresponding simulated confounders. Hence, moving to the right across each row, skill has a greater influence on the selection into treatment (keeping the outcome effect fixed). On the contrary, moving down each column, skill has a greater influence on the untreated outcome (keeping the selection effect fixed).

What Table IV shows is that both the outcome and the selection effect need to be very strong in order to 'kill' the ATT, i.e., to explain almost entirely the positive baseline estimate of the ATT. For low values of the outcome effect, such as $d = 0.1(\Gamma \in [1.6, 1.9])$ in the first row, the point estimate obtained when U is included in the matching set is never smaller than 0.11, and loses its significance only for very high (and quite implausible) values of the selection effect. A comparison with the results of Table III reveals that the cases in which skill is calibrated to match particular observed characteristics of subjects correspond to cells close to the top left of Table IV, with both d and s smaller than 0.2. Thus, the comparison between the two tables suggests that even if the unobserved confounding factor had outcome and selection effects substantially larger than those of the observed covariates, it would not cause much change in the estimated ATT.

While for Tuscany the sensitivity analysis that we have just described conveys an impression of robustness of the matching estimate with respect to reasonable failures of the CIA, a different

[28] Note that the probabilities $Pr(Y = i|T = j)$ and $Pr(T = j)$, with $i, j \in \{0, 1\}$, can be replaced by their sample analogues.
[29] Qualitatively similar results can be derived with different baseline values of $Pr(U = 1)$ and $p_{11.} - p_{10.}$.

DOI: 10.1002/jae

picture emerges in Tables V and VI for Sicily. The baseline estimate for this region indicates an ATT equal to 10 percentage points. However, the sensitivity analysis shows that, for configurations of the parameters that mimic the distribution of important covariates, the baseline result is always 'killed' by the inclusion of U in the matching set. For instance, in the first simulation, the presence of a confounder distributed as gender brings the ATT down to 0.00 (SE = 0.07).

This result is not simply due to the fact that in Sicily observable covariates have a stronger association with selection into treatment, leading us to generate confounding factors U that are more influential than those considered in the case of Tuscany. In fact, in Table VI, the same grid of simulations performed in Table IV for Tuscany shows a very different picture. As soon as U is allowed to have an effect on selection into treatment such that $s = 0.1(\Lambda \in [1.4, 1.6])$, the estimated ATT is halved. Moreover, when U is calibrated to have increasingly stronger selection and outcome effects, the estimated ATT approaches zero at a very rapid pace.

To sum up, while in the absence of the sensitivity analysis one could have argued in favour of a positive effect of TWA employment also in Sicily, the simulations described in Tables V and VI reveal that in this region the estimates are clearly not robust to even minor deviations from the CIA. This finding has a possible explanation. In this region the public sector is the primary source of stable positions and this sector does not recruit through TWAs. The private sector is instead relatively weak and sensitive to business cycle fluctuations. In this context, it is plausible that private firms use temporary workers only as a buffer in order to meet their flexibility needs in the short run, while TWA assignments do not help those who receive them to enter the public sector. Transitions to permanent positions in such a sector are largely dependent on a selection process for which we cannot fully control, and this might explain the lack of robustness of the results in Sicily.

Table V. Sensitivity analysis in Sicily: effect of 'calibrated' confounders

	Fraction $U = 1$ by treatment/outcome				Outcome effect	Selection effect	ATT	SE
	p_{11}	p_{10}	p_{01}	p_{00}	Γ	Λ		
No confounder	0.00	0.00	0.00	0.00	—	—	0.10	0.05
Neutral confounder	0.50	0.50	0.50	0.50	1.00	1.00	0.07	0.06
Confounder-like								
Male	0.87	0.61	0.45	0.27	2.4	5.3	0.00	0.07
Single	0.87	0.82	0.59	0.47	1.7	5.7	0.03	0.07
High school	0.74	0.72	0.76	0.63	2.1	1.5	0.06	0.06
University	0.11	0.06	0.16	0.12	1.5	0.5	0.07	0.06
Prev. employed	0.48	0.30	0.62	0.25	5.4	1.1	0.06	0.06
Prev. permanent	0.07	0.04	0.42	0.06	13.2	0.3	0.09	0.06
Manufacturing	0.17	0.13	0.12	0.03	4.5	3.6	0.04	0.06
Father educ.	0.24	0.25	0.27	0.18	1.8	1.4	0.06	0.06
High distance	0.31	0.26	0.48	0.57	0.7	0.3	0.03	0.06

Note: Let U be a binary confounding factor and denote the fraction of $U = 1$ by treatment and outcomes as: $p_{ij} = Pr(U = 1|T = i, Y = j)$, with $i, j = \{0, 1\}$. On the basis of these parameters, a value of U is imputed to each individual and the ATT is estimated by nearest neighbour propensity score matching with U in the set of matching variables. The process is repeated 1000 times. Γ is the average estimated odds ratio of U in the logit model of $Pr(Y = 1|T = 0, U, W)$; Λ is the average estimated odds ratio of U in the logit model of $Pr(T = 1|U, W)$; 'ATT' is the average of the simulated ATTs; 'SE' is the standard error (calculated as shown in equation (11)). The first two rows show the ATT estimate with no confounding factor or with a confounder whose outcome and selection effects are insignificant, respectively. In the 'confounder-like' rows, U has been calibrated to match the distribution of the corresponding covariate.

DOI: 10.1002/jae

Table VI. Sensitivity analysis in Sicily: characterizing 'killer' confounders

	$s = 0.1$ $\Lambda \in$ [1.4, 1.6]	$s = 0.2$ $\Lambda \in$ [2.2, 2.6]	$s = 0.3$ $\Lambda \in$ [3.6, 4]	$s = 0.4$ $\Lambda \in$ [5.8, 6.5]	$s = 0.5$ $\Lambda \in$ [9.5, 11]	$s = 0.6$ $\Lambda \in$ [17.2, 20]	$s = 0.7$ $\Lambda \in$ [35.7, 44.5]
$d = 0.1\,\Gamma \in [1.7, 2.1]$	0.05	0.03	0.02	0.01	0.00	−0.02	−0.03
	(0.06)	(0.06)	(0.07)	(0.07)	(0.09)	(0.10)	(0.12)
$d = 0.2\,\Gamma \in [2.7, 3.7]$	0.05	0.02	0.01	−0.01	−0.04	−0.07	−0.10
	(0.06)	(0.06)	(0.07)	(0.08)	(0.09)	(0.10)	(0.13)
$d = 0.3\,\Gamma \in [4.2, 6.1]$	0.04	0.02	0.00	−0.03	−0.07	−0.11	−0.16
	(0.06)	(0.06)	(0.07)	(0.08)	(0.09)	(0.11)	(0.12)
$d = 0.4\,\Gamma \in [6.7, 10]$	0.04	0.01	−0.02	−0.06	−0.10	−0.15	−0.21
	(0.06)	(0.06)	(0.07)	(0.08)	(0.09)	(0.10)	(0.12)
$d = 0.5\,\Gamma \in [11.3, 19]$	0.04	0.00	−0.03	−0.08	−0.14	−0.20	−0.28
	(0.06)	(0.07)	(0.07)	(0.08)	(0.09)	(0.10)	(0.12)
$d = 0.6\,\Gamma \in [22, 34.5]$	0.03	0.00	−0.05	−0.10	−0.16	−0.24	−0.32
	(0.06)	(0.07)	(0.07)	(0.08)	(0.09)	(0.10)	(0.12)
$d = 0.7\,\Gamma \in [56.8, 71.6]$	0.03	−0.01	−0.06	−0.12	−0.19	−0.27	−0.37
	(0.06)	(0.07)	(0.07)	(0.08)	(0.09)	(0.10)	(0.11)

Note: Under the assumption that $Pr(U = 1) = 0.4$ and $p_{11} - p_{10} = 0$, the differences $d = p_{01} - p_{00}$ (which captures the outcome effect of U in the absence of treatment) and $s = p_{1\cdot} - p_{0\cdot}$ (which captures the effect of U on the selection into treatment) uniquely define the parameters p_{ij}, with $i, j = \{0, 1\}$. In each cell, the simulated ATT associated to the corresponding differences is reported (standard errors in parentheses). All ATTs are averaged over 1000 iterations. Γ is the average estimated odds ratio of U in the logit model of $Pr(Y = 1|T = 0, U, W)$; Λ is the average estimated odds ratio of U in the logit model of $Pr(T = 1|U, W)$. The baseline estimate without confounder is equal to 0.10 in Sicily. The Manski bounds are (−0.76, 0.24).

5. CONCLUSIONS

The diffusion of TWA jobs originated a harsh policy debate and ambiguous empirical evidence. Results based on quasi-experimental evidence (uniquely coming from US data) suggest that a TWA assignment decreases the probability of finding a stable job, while results based on the CIA (mostly coming from European data) reach opposite conclusions. Using data from two Italian regions, specifically collected for this evaluation study, we use a matching estimator to show that TWA assignments may be an effective springboard to permanent employment. We also propose a sensitivity analysis for matching estimators, which in our empirical study highlights that only for one of the two regions, Tuscany, are the results robust to specific failures of the CIA.

We conclude that non-experimental studies on the effects of TWA employment (i.e., studies based on the CIA) should not be automatically discarded because they lack exogenous variation in assignment to treatment. They should, however, be put under the scrutiny of a sensitivity analysis like the one we propose before being accepted as a guide for policy. This conclusion is relevant for the debate originated by the opposite findings on the effects of TWA jobs in Europe and in the USA. Inasmuch as the European results could be shown to be robust to failures of the CIA with a sensitivity analysis like the one we propose, the lack of a quasi-experimental basis would not be a sufficient reason to discard them.

This line of argument is even more compelling given that there are institutional reasons to expect different effects of TWA jobs on the two sides of the Atlantic. For example, firing costs are lower in the USA than in all the European countries where the effect of TWA employment has been evaluated. The higher are firing costs for stable contracts, the larger the scope for TWA jobs as a screening device, since firms attribute greater importance to the assessment of the quality

DOI: 10.1002/jae

of workers before locking themselves into a new employment relationship. In this context, for most workers the availability of TWA assignments increases the probability of a transition to a permanent job. At the same time, higher firing costs may induce firms to use temporary workers as a mere flexibility buffer, if they make it impossible to adjust the number of regular employees during business cycle downturns. If the first effect dominates the second, one should observe a stronger springboard effect of TWA employment where firing costs are higher. If, on the contrary, the second effect is the prevailing one, the springboard effect should be weaker where firing costs are higher. Since we cannot say which effect dominates only on theoretical grounds, we may very well expect different springboard effects of TWA jobs in countries with different employment protection regimes.

ACKNOWLEDGEMENTS

Financial support by the Italian Ministry of Welfare and the Tuscany Region is gratefully acknowledged. We also thank Manpower Italia for help in data collection, and seminar participants at AIEL, CEMFI, Perugia, Oviedo, the 2004 CNR meeting in Venice, the EC workshop on 'Temporary Work in Europe', the 2004 CEA Annual Meeting, the 2004 ZEW Conference on Evaluation Research, and the 2005 Italian Congress of Econometrics for insightful comments and suggestions. A previous version of this paper has circulated with the title 'Sensitivity of matching estimators to unconfoundedness: an application to the effect of temporary work on future employment'. A routine program that implements the proposed sensitivity analysis in Stata (see Nannicini, 2007) can be downloaded at www.tommasonannicini.eu, or by typing 'net search sensitivity matching' within Stata.

REFERENCES

Abadie A, Imbens GW. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**: 235–267.

Altonji JG, Elder TE, Taber CR. 2005. Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools. *Journal of Political Economy* **113**: 151–184.

Amuedo-Dorantes C, Malo MA, Munoz-Bullon F. 2006. The role of temporary help agencies in facilitating temp-to-perm transitions. IZA Discussion Paper 2177.

Anderson P, Wadensjö E. 2004. Temporary employment agencies: a route for immigrants to enter the labour market? Discussion Paper 1090, IZA.

Autor DH, Houseman S. 2005. Do temporary help jobs improve labor market outcomes for low-skilled workers? Evidence from random assignments. Mimeo, MIT.

Black D, Smith J. 2004. How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of Econometrics* **121**: 99–124.

Booth AL, Francesconi M, Frank J. 2002. Temporary Jobs: stepping stones or dead ends? *Economic Journal* **112**: 189–213.

Caliendo M, Kopeinig S. 2007. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*.

Dehejia RH, Wahba S. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics* **84**: 151–161.

European Commission. 2003. *Employment in Europe*. European Commission: Brussels.

Gerfin M, Lechner M, Stieger H. 2002. Does subsidized temporary employment get the unemployed back to work? An econometric analysis of two different schemes. CEPR Discussion Paper 3669.

Heckman JJ, Ichimura H, Todd P. 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Review of Economic Studies* **64**: 605–654.

DOI: 10.1002/jae

Heckman JJ, Todd P. 1999. Adapting propensity score matching and selection models to choice-based samples. Mimeo, University of Chicago.
Ichino A, Mealli F, Nannicini T. 2005. Temporary work agencies in Italy: a springboard toward permanent employment? *Giornale degli Economisti e Annali di Economia* **64**: 1–27.
Ichino A, Mealli F, Nannicini T. 2006. From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity? CEPR Discussion Paper 5736.
Imbens GW. 2003. Sensitivity to exogeneity assumptions in program evaluation. *AEA Papers and Proceedings* **93**: 126–132.
Imbens GW. 2004. Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* **86**: 4–29.
Kvasnicka M. 2005. Does temporary agency work provide a stepping stone to regular employment? SFB Discussion Paper 649, Humboldt University.
Lane J, Mikelson KS, Sharkey P, Wissoker D. 2003. Pathways to work for low-income workers: the effect of work in the temporary help industry. *Journal of Policy Analysis and Management* **22**: 581–598.
Lechner M. 2002. Program heterogeneity and propensity score matching: an application to the evaluation of active labor market policies. *Review of Economics and Statistics* **84**: 205–220.
Lechner M, Pfeiffer F, Spengler H, Almus M. 2000. The impact of non-profit temping agencies on individual labour market success. ZEW Discussion Paper 00–02.
Malo MA, Munoz-Bullon F. 2002. Temporary help agencies and the labour market biography: a sequence-oriented approach. FEDEA, EEE 132.
Manski CF. 1990. Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings* **80**: 319–323.
Michalopoulos C, Bloom HS, Hill CJ. 2004. Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *Review of Economics and Statistics* **86**: 156–179.
Nannicini T. 2004. The take-off of temporary help employment in the Italian labor market. EUI-ECO Working Paper 09/04.
Nannicini T. 2007. Simulation-based sensitivity analysis for matching estimators. *Stata Journal* **7**(3): 334–350.
OECD. 2002. Taking the measure of temporary employment. In *Employment Outlook*. OECD: Paris. pp.135–196.
Rosenbaum P. 1987. Sensitivity analysis to certain permutation inferences in matched observational studies. *Biometrika* **74**: 13–26.
Rosenbaum P, Rubin D. 1983a. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B* **45**: 212–218.
Rosenbaum P, Rubin D. 1983b. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**: 41–55.
Rubin D. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York.
Smith J, Todd P. 2005. Does matching overcome Lalonde's critique of nonexperimental estimators? *Journal of Econometrics* **125**: 305–353.
Wang L, Krieger AM. 2005. Causal conclusions are most sensitive to unobserved binary covariates. *Statistics in Medicine* **25**: 2257–2271.

DOI: 10.1002/jae