STATISTICA, anno LXXI, n. 2, 2011

FIFTEEN YEARS OF LABOUR MARKET REGULATIONS AND POLICIES IN ITALY: WHAT HAVE WE LEARNED FROM THEIR EVALUATION?*

Ugo Trivellato

1. BACKGROUND AND MOTIVATION

During the last fifteen years a notable number of labour market interventions took place in Italy, under two main headings: new regulations, occasionally coupled with reforms of institutions; implementation of new – and reforms of previous – active and/or passive programmes. Stimulated also by advances in programme evaluation methods, there has been a growth of empirical studies aimed at estimating the effects of these interventions.

The comparative analysis of impact evaluations of labour market policies (LMPs), regulations included, is still in its infancy. LMPs are diversified, country and context specific. They hardly meet protocols that allow a sensible aggregation of evidence across studies in ways that yield statistically rigorous and readily meaningful estimates of effectiveness[1]. Thus, most of the meta-analyses on the effects of LMPs take on a narrative style[2].

Hitches are particularly severe for the case of Italian LMPs. They were not designed with a concern for the assessment of their impact; as a consequence, all the evaluation studies are observational. Besides, the effects of a given policy have been investigated by a limited number of studies. Moreover, they are rather idiosyncratic with respect to various characteristics (features the policy, area

* This is a substantially revised and abridged version of the Keynote Address at the AIEL, LABORAtorio Riccardo Revelli and CHILD International Conference on Labour Market and the Household (November 6-7, 2009, Fondazione Carlo Alberto, Moncalieri, Italy). I am indebted to conference participants for useful criticisms and comments and to my colleagues at IRVAPP – chiefly Erich Battistin, Adriano Paggiaro, Enrico Rettore and Antonio Schizzerotto – for joint work and many discussions on the topic. The usual disclaimer applies.

[1] By contrast, a well established system of research synthesis exists in the area of education: see the What Works Clearinghouse [http://ies.ed.gov/ncee/wwc/] and the Campbell Collaboration on education, crime and justice [www.campbellcollaboration.org/]. The standard is provided by the systematic review of the effects of health care carried out by the Cochrane Collaboration [www.cochrane.org/].

[2] A recent exception is Card, Kluve &Weber (2009).

and/or population of interest, data sources, evaluation design, *etc.*). Thus, a rigorous synthesis is definitely problematic.

The purpose of this paper is modest. After a cursory outline of the LMPs implemented in Italy from the early '90s, I briefly summarize the evidence from some subjectively selected studies published in the last ten years (Section 3).

Then, I focus on two aspects. First, I look at a blend of empirical and analytical issues critical for credible impact evaluations, that emerge from such review. They refer to prospective *vs.* retrospective evaluation, availability (or lack) of adequate data, over-identification tests in order to corroborate (or falsify) the identifying restriction on which the evaluation method rests, and heterogeneous effects (Section 4).

Second, I survey the substantive evidence offered by the reviewed studies. It points to generally minor policy effects, and I present some tentative explanations for that lack of effectiveness (Section 5).

Preliminarily, to make the exposition self-contained, I sum up (less than) the barebones of counterfactual analysis, the framework for causal inference developed mainly by Donald B. Rubin & co-authors and James J. Heckman & co-authors[3] (Section 2).

2. COUNTERFACTUAL ANALYSIS IN A NUTSHELL

2.1. *The selection bias problem*

Consider a well defined policy (≡ intervention, programme, treatment) targeted to a well defined population, whose members can in principle be (self-)assigned to/(self-)denied the intervention, with the purpose of inducing an effect on a well defined state or behaviour of the units exposed to the intervention. Let that state or behaviour be measured by an outcome variable, Y say. For simplicity, let the policy consist of a single treatment, with the binary variable D denoting the treatment/non-treatment status.

Let Y^T and Y^{NT} be the potential outcomes a specific unit would experience being exposed to the treatment and denied it, respectively. For each unit the causal effect is logically defined as the difference of the potential outcomes under treatment and no treatment, respectively. But it is apparent that we can observe only one of the potential outcomes for each unit, depending on the treatment status actually experienced by it. This is «*the fundamental evaluation problem*» (Heckman, Lalonde & Smith, 1999, p. 1879), or indeed, more broadly, «*the fundamental problem of causal inference*» (Holland, 1986, p. 947).

A sensible approach consist of focusing on the identification of causal parameters, *i.e.*, specific features of the distribution of the causal effect $Y^T - Y^{NT}$ for (a subset of) the reference population. The identification problem amounts to de-

[3] The literature of the topic is enormous. For recent reviews see, *e.g.*, Blundell & Costa Dias (2009) and Imbens & Wooldridge (2009).

rive conditions under which a causal parameter can be recovered from the distribution probability of suitable observed variables.

Typically, an interesting causal parameter is a mean. For simplicity, let us consider the average treatment effect on the treated (ATT):

$$ATT = E\,[a \mid D{=}1] = E\,[Y^T - Y^{NT} \mid D{=}1] = E\,[Y^T \mid D{=}1] - E\,[Y^{NT} \mid D{=}1] \quad (1)$$

The last term in equation (1) is a counterfactual, unobservable by construction, since the outcome Y^{NT} is never observed on those undergoing treatment. We do observe the mean value of Y^{NT}, but only on the non-treated group. By contrasting it to the mean outcome experienced by the treated group we get the following identity:

$$E\,[Y^T \mid D{=}1] - E[Y^{NT} \mid D{=}0] = E\,[a \mid D{=}1] + (E\,[Y^{NT} \mid D{=}1] - E[Y^{NT} \mid D{=}0]) \quad (2)$$

Equation (2) clarifies that the observed mean difference between treated and non-treated includes the selection bias, namely, the difference we would have observed had the treated been denied the treatment. It arises when participation depends on characteristics that do affect Y^{NT} and are unequally distributed between treated and non-treated. Note also that the selection bias term too involves a counterfactual outcome.

Patently, there is no unique recipe for a way out from the selection bias problem. One should try to understand as much as s/he can about the selection process, in order to adequately specify it.

Let D, the binary treatment status, be a deterministic function of a realized value of the triple $\{X, U, Z\}$, where:

– X is a set of observable characteristics of the units, unaffected by the intervention (typically measured prior to it), possibly correlated to the outcome Y^{NT};

– U are unobservable characteristics of the units, unaffected by the intervention, possibly correlated to the outcome Y^{NT};

– Z is the observable binary outcome of a random draw, thus independent of the potential outcomes.

Then $D(X, U, Z)$ properly represents the selection process, *i.e.*, a mapping from the space $\{X, U, Z\}$ onto the space $\{1,0\}$.

2.2. *A stylized taxonomy of special cases of the selection process*

In order to provide a convenient frame for the subsequent review of the evaluation studies, it is useful to characterize some main patterns of the selection process which emerge as special cases of $D(X, U, Z)$, along with the corresponding strategies for identifying average treatment effects. I move from the "ideal" case of a randomized experiment and then focus on policies implemented in an observational setting.

Randomized experiment: $D(X, U, Z) = Z$.

Under randomization, selection bias vanishes by construction. $E\,[Y^T \mid D{=}1] - E[Y^{NT} \mid D{=}0]$, the mean difference between the treated and non-treated group, straightforwardly identifies the average treatment effect.

Thus, the randomized experiment is a sort of benchmark, against which one has to assess the properties of the other identification strategies. All these strategies aim at mimicking the fundamental feature of an experimental design: having two groups equivalent in all relevant respects – the *ceteris paribus* clause – but different with respect to the probability of being exposed to the intervention.

When implemented for policies targeted to agents – individuals, households, firms, *etc.* –, randomization typically induces behavioural responses from treated and/or non-treated units. Thus, randomized experiments should be regarded as «*the bronze standard*», rather than the gold standard (Berk, 2005).

Selection on observables: $D(X, U, Z) = D(X, U)$, with U independent of the potential outcomes.

The enforcement of the *ceteris paribus* clause only requires conditioning on X[4]. As a result the composition of the two groups is made equivalent with respect to X.

The rationale for this strategy rests upon the claim that all differences between treated and non-treated units relevant to the outcome variable that enter the selection process are captured by the observable variables X. Once all these – and only these – factors are controlled for in the analysis, the selection bias term is zero by definition and thus the average treatment effect can be retrieved. Stratification and matching, especially propensity score (p-score) matching – in order to avoid dimensionality problems –, emerge as appropriate.

Regression Discontinuity Design (RDD): $D(X, U, Z) = D(X)$.

This is typically the case when the selection process is driven by administrative rules. For simplicity, let X be a scalar with $D=\mathrm{I}(X>x_0)$, where x_0 is a known point in the support of X, and I is an indicator variable taking the value 1 when the expression within brackets is true[5]. This set-up defines a (sharp) RDD[6].

ATT is identified locally, around x_0, by comparing units just on the right of x_0 to units just on the left of it. Indeed, these two groups are approximately equivalent with respect to X, the only variable driving the selection process. But «*the design has fundamentally only a limited degree of external validity, although the specific average effect that is identified may well be of special interest*» (Imbens & Lemieux, 2008, p. 628).

Difference-in-differences design: $D(U, Z) = D(U)$, with U and Y^{NT} dependent, but U independent of the variation of Y^{NT} over time.

[4] The two groups might not be equivalent with respect to U, but U is assumed to be independent of the potential outcomes.

[5] It is worth noting that this case is different from the previous one, because conditional on $X=x_0$ the two groups are equivalent with respect to any other characteristic, whether observable or not.

[6] The qualification "sharp", dropped in the sequel, points to the fact that the treatment status deterministically follows from the rule $D=\mathrm{I}(X>x_0)$. By contrast, we have a "fuzzy" RDD when individuals are first ranked according to an observable indicator X and then assigned to the intervention according to the rule $Z=\mathrm{I}(X>x_0)$, but because of non-compliance the actual treatment status D is different from Z.

Under the assumed selection process, this design is appropriate when Y is a repeatable event (wages, say), observed both before and after the intervention. Thus, the minimal information set consists of a pair of observations – the means before and after the intervention – for the treated and non-treated group, respectively.

ATT is identified by the difference of the average variation over time of the treated and non-treated group, respectively[7].

Natural experiment: $D(X, U, Z) = D(U, Z)$.

The term "natural experiments" is used to refer to «*situations where the forces of nature or government policy have conspired to produce an environment somewhat akin to a randomized experiment*» (Angrist & Krueger, 2001, p. 73). Here $D \neq Z$ because of non-compliance of some units to the assignment, and $E[Y^T | D=1] - E[Y^{NT} | D=0]$ is affected by selection bias since U is correlated to Y^{NT}. Note, however, that by construction the two groups indexed by Z are equivalent with respect to any characteristic relevant for the potential outcomes.

The so-called "intention to treat", *i.e.*, the causal effect of being assigned to the treatment, is easily identified as $E[Y | Z=1] - E[Y | Z=0]$.

More conditions are needed to identify a meaningful average causal effect of the intervention using Z as an instrumental variable (IV) for D.

2.3. *Testable implications and over-identification tests*

The evaluation strategies just outlined rely on the analyst's ability to understand the selection process. This knowledge translates into an identifying restriction, which we can express in the form of a conditional independence assumption (CIA)[8]. For instance, in the case of selection on observables, one assumes $Y^{NT} \perp D | X$.

For a credible identification of the effects in an observational setting, it is important to check such assumption. This amounts to derive implications of the identification strategy exploited that can be tested against data. They will provide necessary conditions for the internal validity (in the sense that they might mistakenly fail to refuse it) of the estimated causal parameters, and can thus be used to corroborate causal conclusions.

This route needs supplementary information, either in form of additional data or of "qualitative" information on the way in which the treatment produces its effects. (a «*causal mechanism or theory*» in the sense of Rosenbaum, 1984, p. 43). If data are against the conjunction of this supplementary information and the identi-

[7] A straightforward generalization, quite important for applications, arises when $D(X, U, Z) = D(X, U)$, with U and Y^{NT} dependent, but U independent of the variation of Y^{NT} over time. Y is still assumed to be a repeatable event, which is observed both before and after the intervention. Here we face both selection on observables and the peculiar dependence of U and Y^{NT} of the case just above. A sensible strategy is in two steps: (i) conditioning on X; (ii) then resorting to difference-in differences.

[8] Or ignorable treatment assignment, in Rubin's terminology.

fying restriction imposed, this casts doubts on the identification strategy employed.

The implementation of over-identification tests sensibly relies on some general guidelines. They include tests involving variables that are, by definition, not affected by the intervention (*e.g.*, variables measured prior to treatment assignment – and unaffected by expectations about it –, multiple comparison groups), and specification testing procedures for selecting an appropriate non-experimental estimator (Heckman & Hotz, 1989). However, it clearly depends on the evaluation strategy adopted. Over-identification tests of general applicability have been suggested for specific selection processes[9]. Still, in several instances over-identification tests should be tailored to the specific traits of the evaluation design and rest upon the peculiar supplementary information available.

3. A CURSORY REVIEW OF FIFTEEN YEARS OF LMPS AND EVALUATION STUDIES

3.1. *LMPs in Italy from the early '90s: A concise outline*

The picture of recent labour market regulations and policies in Italy is quite intricate, *per se* and because it impinges upon a fragmented system. Useful presentations are in Sestito (2002), Pirrone & Sestito (2006; 2009), Anastasia, Mancini & Trivellato (2009), and Anastasia, Paggiaro & Trivellato (2011). I just sketch the main novelties on LMPs brought in from the early '90s to 2007.

- 1990-94-97-99: the *Contratto di formazione e lavoro* (CFL) is modified. CFL is a fixed-term – 1 to 2 years – training and labour contract for young people, with lower labour costs, no firing costs and indirect incentives to end up in an open-ended contract. Modifications include region- and industry-variations.
- 1991 and 1993: the programme *Liste di mobilità* (LiMo) is introduced. It is targeted to dismissed workers, and combines passive and active measures – income support and incentives to firms for re-hiring workers, respectively – that vary deterministically with worker's age (<40, 40-49, ≥50 years) and firm's size (15 employees threshold).
- 1997: the so-called "Treu reform" is approved. Its main provisions are the introduction of Temporary Work Agency (TWA) employment and an initial reform of Public Employment Services (PES).
- 2001: the use of fixed-term contracts is made much easier. Basically, no justification is required for them.
- 2001-05-07: three modifications of the ordinary unemployment insurance scheme are brought in, with progressively higher replacement ratio and longer maximum duration.
- 2003: the so-called "Biagi Act" is approved. It extends the opportunities for using apprenticeship, introduces the *contratto d'inserimento* (worker's fist job con-

[9] For example, for selection on observables see Rosenbaum (1984, 1987); for RDD see Imbens & Lemieux (2008), Lee (2008), McCrary (2008).

tract) as well as a set of flexible contracts (job-on-call, job-sharing, *etc.*), completes the reform of PES.

In addition, in 2002 there was an unsuccessful attempt to modify art. 18 of the *Statuto dei lavoratori* (the basic act on workers' rights, dating back to 1970), that sets a stringent employment protection against unfair individual dismissal of workers by firms over 15 employees. The issue is worth mentioning, because it stimulated research on the effect of art. 18 on the firms' propensity to grow around the 15 employees' threshold.

Overall, these interventions have been qualified as marginal – they apply to new entrants (and re-entrants) only – and incomplete – fundamentally they deal with new labour contracts, with poor attention paid to the reform of the welfare – (Sestito, 2002, p. 17). If one places there interventions in the context of the LMPs adopted by most Western European countries (see, *e.g.*, OECD, various years; Grubb & Martin, 2001; Bassanini, Nunziata & Venn, 2008; Martin & Scarpetta, 2010;), a fairly neat polarization emerges. In Italy, the strategies favouring flexibility (OECD, 1994) have been taken into account, while the subsequent shift of OECD' suggestions and the indications of the Commission of the European Communities (2007) towards flexicurity have been *de facto* ignored.

3.2. *A synopsis of selected evaluation studies*

A review of evaluation studies of Italian LMPs is in Trivellato & Zec (2008)[10]. Drawing from it and extending consideration to subsequent contributions, I selected seventeen studies, that cover various LMPs policies and present a variety of empirical strategies for assessing the effect of interventions. A synopsis of these studies in Table 1.

The first group of studies deals with labour market regulations. The first two studies pertain to the effect of art. 18 of the *Statuto dei lavoratori*. The other seven studies focus on flexible contracts: Ichino, Mealli & Nannicini (2005) on TWA employment; Barbieri Scherer (2007) on apprenticeship, contrasted to temporary (≡ fixed-term) contracts; Gagliarducci (2005), Barbieri & Sestito (2008) and Paggiaro, Rettore & Trivellato (2010) on temporary contracts; Berton, Pacelli & Devicienti (2008) and Bison, Rettore & Schizzerotto (2010) on a variety of flexible contracts.

The only appreciable intervention concerning the institutional setting is the reform of PES. Two studies aimed at assessing its effects are examined.

The last six studies consider LMPs *stricto sensu*: two deal with vocational training programmes, one with the CFL, and three with the LiMo programme.

Table 1 is self-explanatory. I do not comment on the single papers. Instead, I consider a set of empirical, analytical and substantive issues relevant for evaluations carried out in an observational setting – which is the case for all reviewed studies, and look at the evidence that these studies –, provide on them.

[10] I do not consider incentives to firms with an employment target. A review of evaluation studies of policies consisting of firms' incentives is in Ercoli & Guelfi (2008).

TABLE 1

A synopsis of selected studies

Reference	*Programme/ Intervention*	*Outcomes*	*Area, time & data*	*Evaluation method*	*Identification issues*(a)	*Causal effects*
Regulations						
Garibaldi, Pacelli & Borgarello (2004)	Art.18: more stringent EPL for firms > 15 employees.	Firms' propensity to grow around the threshold.	Italy,1987-96. WHIP(b), firms ≤ 30 employees.	RDD-type (+ model).	Implied by the model. Rating: ordinary.	Significant, but quantitatively small.
Schivardi & Torrini (2008)	Art.18: more stringent EPL for firms > 15 employees.	Firms' propensity to grow around the threshold.	Italy, 1986-98. Full INPS(c) data set, firms with 5-25 employees.	RDD (+ reduced-form model).	Assumed (supported by the data). Rating: decent.	As above. Firms >15 have less stable employment relations.
Ichino, Mealli & Nannicini (2005)	Temporary Work Agency (TWA) employment *vs.* unemployment or other atypical jobs.	Permanent employment 18 months later.	Tuscany & Sicily, selected provinces, 2001.1. One TWA's data & *ad hoc* phone survey.	p-score matching (provinces with/ without TWAs).	CIA questioned. Sensitivity analysis for potential confounders. Rating: ordinary.	Positive effect of TWA employment just *vs.* unemployment just in Tuscany.
Barbieri & Scherer (2007)	Apprenticeship *vs.* temporary contracts (flows into).	Employed, Permanently employed 2, 3, 4 years later.	Veneto, 1998.8-99.7. Giove(d).	p-score matching.	Balancing property. Rating: decent.	Positive effects of apprenticeship.
Gagliarducci (2005)	Temporary contracts (flows into), chiefly repeated temporary contracts.	Permanent employment.	Italy, people aged 18-55, 1997. ILFI(e), 1997 wave & retrospective information.	Multi-spell proportional hazard model with competing risks.	No, but flexible specification: nonparametric baseline & unobs. heterogeneity. Rating: ordinary.	Repeated temporary jobs with interruptions decrease the hazard to a permanent job.
Berton, Pacelli & Devicienti (2008)	4 types of temporary contracts, in a framework of 6 employment states + "non employment".	Yearly transition probabilities.	Italy, new entrants aged 15-39, 1998-2004. WHIP(b).	Markov chain model with fixed effects.	Markov chain assumed. Rating: questionable.	"Port-of-entry" effects, especially for training contracts & within-firm.
Barbieri & Sestito (2008)	Temporary contracts *vs.* unemployment (flows into), in three different years (≡ regulations).	Employed, Permanently employed., Satisfactorily employed 1 year later.	Italy, 1993, 1999 and 2002. QLFS(f).	p-score matching.	Balancing property (dubious). Rating: questionable.	Significant positive effect on Employed & Satisfactorily employed.
Bison, Rettore & Schizzerotto (2010)	"Treu reform" *vs.* pre-reform regulations.	7 monthly labour force states, moonlight employment included, over 3 years.	Italy, new entrants in 1993-95 (pre-) and 1999-2001 (post-reform). ILFI(e).	p-score matching.	*Placebo* test for CIA (applied to two cohorts subject to the same regulations). Rating: High.	Moderate positive effects of the "Treu reform" after 3 years: more employed, less moonlight employed, less unemployed.
Paggiaro, Rettore & Trivellato (2010)	Temporary contracts *vs.* unemployment (flows into), in three different periods (≡ regulations).	Employed, Permanently employed, Satisfactorily employed 1 year later.	Italy, three periods: 1995-96, 2000-01, 2005-06. Q/CLFS(f), eight 4-wave panels from for each period.	p-score matching.	Backward test for CIA (see main text, Section 4.5). Credible results for 2005-06 only. Rating: High.	For 2005-06: Positive effects for Employed and Satisfactorily employed. North & Centre *vs.*-South heterogeneity for men.

TABLE 1

follows (I)

Reference	*Programme/ Intervention*	*Outcomes*	*Area, time & data*	*Evaluation method*	*Identification issues*(a)	*Causal effects*
Institutions						
Barbieri G. *et al.* (2003)	Public Employment Services (PES) at provincial level: enrolled/not enrolled to PES (4 types) and indicator of the quality of PES.	Employed 3 months later.	Italy, 2001. QLFS(f), 2-wave panel.	p-score matching. IV.	Balancing property. Curse of dimensionality Rating: ordinary.	Basically no effects of PES.
Naticchioni & Loriga (2008)	Unemployed at t_1 enrolled in the PES in the subsequent quarter *vs.* unemployed at t_1 not enrolled in the PES in the subsequent quarter.	Employed within days and 9-12 months later. Permanently employed as above.	Italy, 2004.I-2006:III. QLFS(f), six 4-wave panels.	p-score matching.	Robustness checks (slight changes in the groups). Sensitivity analysis for potential confounders. Rating: decent.	Employed: negative effect in the short-, positive in the long-term. Permanently employed: no effect. PES less effective in the South.
Active and/or passive LMPs						
Battistin & Rettore (2002)	Vocational training programme: participants *vs.* applicants excluded (based on a test score).	Employed 6 months later.	Turin, 1995.10-96.6 *Ad hoc* survey.	Fuzzy RDD (non-compliers among the controls).	Rating: high.	No effect (from 2 tests for discontinuity in fuzzy RDD).
Berliri, Bulgarelli & Pappalardo (2002)	Participants to European Social Fund co-financed vocational training programmes *vs.* first-time unemployed with no training.	Employed in 1998.	Lombardy & Emilia, 1997 *Ad hoc* survey + panel from the QLFS(f).	IV-type estimator: bivariate probit & selection equation.	Rating: poor.	Positive, more for men and participants with high school diploma.
Contini *et al.* (2002)	Training-labour contracts (CFL) in the private sector, with variations of labour and firing costs –over time and across areas – for an eligible worker relative to a non eligible one, due to various reforms.	Employment state: during the eligibility period; after the eligibility period.	Italy, 1986-1996. WHIP(b), cohorts born in 1958-77, tracked over the age window 19-34.	Regression-type model that chiefly exploits the variations above.	Model specification, IV estimation. Rating: ordinary.	No effect on the chance to get a job during eligibility. Positive effect of work experience with CFL
Caruso & Pisauro (2005)	*Liste di mobilità* (LiMo): programme targeted to dismissed workers, that combines income support to workers and incentives to firms for re-hiring workers. These measures vary deterministically with worker's age and firm's size.	Hazard rate to permanent employment for workers dismissed by large firms: differential age effects.	Umbria, 1995.1-98.12. Linkage of LiMo archive & Netlabor(d).	Semi-parametric Cox's model. Competing risks model for recalls (same firm) & new jobs.	Selection on observables *via* a model. Rating: questionable.	The hazard rate declines with the length of the eligibility period for the total and recalls, not for new jobs.

TABLE 1

follows (II)

Reference	*Programme/ Intervention*	*Outcomes*	*Area, time & data*	*Evaluation method*	*Identification issues*[a]	*Causal effects*
Martini & Mo Costabella (2007)	LiMo.	Monthly employment rate over 3 years after enrolment: differential age and income support effects.	Turin province (less the main city), 1997-2000. Netlabor[d], 1995-2003.	p-score matching. RDD.	*Placebo* test for CIA within age-groups: 1st group restricted to dismissed workers aged 30-39. Rating: decent for RDD; questionable for matching.	Negative income support effect. Effect of 1-year longer eligibility: no without income support; negative with income support.
Paggiaro, Rettore & Trivellato (2009)	LiMo.	Differential age effects on: - monthly employment state, - gross real weekly wage (2003 prices; conditional on employment), over 3 years after enrolment.	Veneto, 1995-98 Linkage of Netlabor[d] & INPS[c], 1992-2001	RDD.	*Placebo* tests for CIA within age-groups refute matching. Over-identification tests (comparing individuals around the thresholds with respect to their pre-programme work history) validate RDD. Rating: high.	Stratified by gender and eligibility to income support: - at the 40-year threshold (a) no effect for men, (b) older women with income support postpone re-entry at work; - at the 50-year threshold strong negative effect for workers with income support.

[a] This items conclude with a subjective, overall assessment of the credibility of the identifying restriction and over-identification checks (sensitivity analyses, over-identification tests, *etc.*). I formulate the assessment on a five-point scale: high/decent/ordinary/questionable/poor.

[b] INPS is the Italian social security agency.

[c] WHIP is the acronym for Work Histories Italian Panel, an employer-employee micro-data set released by LABORatorio R. Revelli starting from the INPS archives.

[d] Netlabor and Giove are the first- and second-generation data archives, respectively, released by Veneto Lavoro starting from information collect by Labour Exchange Offices.

[e] ILFI is the acronym for *Indagine Longitudinale sullle Famiglie Italiane*, a household panel survey carried out by the Department of Sociology, University of Trento.

[f] Q/CLFS designates the Italian Labour Force Survey, Quartely (up to 2003) and Continuous (from 2004), respectively, carried out by Istat.

4. SOME EMPIRICAL AND ANALYTICAL ISSUES ABOUT IMPACT EVALUATION OF ITALIAN LMPS

4.1. *Prospective vs. retrospective evaluation*

Typically, impact evaluations refer to a policy which has been implemented. However, an important distinction should be made between retrospective evaluations, in the specific sense that they are designed and conducted after the policy has been designed and implemented, and prospective evaluations, which are developed at the same time as the programme is being designed and are built into programme implementation.

In general, a prospective impact evaluation is more likely to produce credible results, for various reasons. First, it stimulates an informed debate, useful for learning about the policy design. Second, it helps focusing on implementation aspects, fist of all on clear and transparent assignment rules, crucial for generating valid counterfactuals. Third, it creates an incentive – and opportunities – to timely assemble the information necessary to assess results[11].

By contrast, retrospective evaluations are dependent on a clear design of the policy, on its consistent implementation, and on the availability of relevant data with sufficient coverage of the treatment and comparison groups over time. Failure to meet this conditions might challenge the very same feasibility of a retrospective evaluation. In any case, a retrospective evaluation has to use identification strategies for the observational setting at hand, and is forced to rely on stronger assumptions. Hence, they can produce evidence that is more debatable.

Unfortunately, all the reviewed evaluations – indeed, as far as I know, all evaluations of Italian LMPs – are retrospective[12]. Thus, it comes with no surprise that sometimes knowledge of the possibly blurry implementation of a policy is inadequate (see Caruso &. Pisauro, 2005, for evidence about the implementation of LiMo in Umbria). Besides, the large majority of the reviewed studies had to use – or adapt – general purpose micro databases.

From a broader perspective, the state of the art reveals a poor interest of the policy-makers and the public opinion to learn about the effectiveness of an intervention. The landscape of Italian LMPs evaluation is almost void of demonstrations, even more of small-scale randomized experiments. To the best of my knowledge, the only exception is Martini (2009): a Randomized Clinical Trial to test a programme to place mentally ill patients into permanent jobs, just set off.

4.2. *The crucial role of appropriate data for sound (LM) policies evaluations*

The role of adequate data for credible impact evaluations has been exemplarily stressed by Heckman, LaLonde & Smith (1999, pp. 1868-1869):

> *Better data help a lot. The data available to most analysts have been exceedingly crude. Too much has been asked to econometric methods to remedy the defects of the underlying data. [...] The best solution to the evaluation problem lies in improving the quality of the data on which evaluations are conducted and not in the development of formal econometric methods to circumvent inadequate data.*

When the analyst is forced to use general purpose micro databases for evaluating the effects of an intervention, typically s/he faces two problems.

[11] The importance of prospective evaluation has been stressed in the so-called Barca Report: «*To facilitate and to make more effective the counterfactual approach, it must be used* prospectively *(designing impact evaluation in tandem with policy design), not retrospectively (designing and conducting it after policy has been designed and implemented). By making explicit the expected results and the linkages between means and ends on which the intervention is based, and by building a strategy for learning about policy effects, impact evaluation can contribute to [...] the clear identification of objectives in policy design. [...] Many of the necessary data could (or can only) be collected in the process of implementing the interventions which they should help evaluating, but they fail to be collected because their need is not identified in time*» (Barca, 2009, pp. 179-180).

[12] This is the case also for the few studies carried out under a contract from the governmental agency responsible for designing the policy, specifically the Ministry of Labour: Ichino, Mealli & Nannicini (2005) and Paggiaro, Rettore & Trivellato (2009).

(a) The target population of the policy is often a small segment of the database's population. If the source of the data is a large purpose sample survey, the portion of the database relevant for impact evaluation is a small, typically unplanned domain. Possibly severe issues of lack of precision – the sample size is too small to detect the effect – and/or potential biases – precisely because the domain is unplanned – come up.

(b) The outcome variable(s) and other variables needed for sensibly controlling for the selection on observables might be not fully available: that is, relevant data are lacking.

The main Italian general purpose micro databases on LMPs face exactly these problems[13]. Let us consider first the databases from administrative archives: INPS, Labour Exchange Offices, and the panel versions resulting from them – WHIP and Netlabor/Giove, respectively – (see Table 1, footnotes (b)-(d)). These are census-type databases, with fairly detailed work histories. They suffer from two key limitations: first, they do not cover the entire working/labour force population, the consequence being that exits from the scope of the archives result in truncated or incomplete work histories; second, they have no information on education – a key variable – and the household.

As for large purpose recurrent sample surveys, a pivotal role is played by the (previously quarterly, from 2004 continuous) Labour Force Survey (Q/CLFS). Among its main advantages, three are noteworthy: it covers the entire population of interest; it collects fairly detailed information on participation at work and search for work; it allows to link individual records of all household's members. By contrast, it does not currently collect information on income and wages, nor to LM programmes; besides, it has a moderate sample size and its rotating sample design allows one to get just short and fragmented panels.

Some evaluation studies utilize household panel surveys. Among those reviewed this is the case for Bison, Rettore & Schizzerotto (2010). Several other studies use the panel component of the Bank of Italy's Survey on Household Income and Wealth (see Trivellato & Zec, 2008). Their main pro consists in the fact that they provide consistent longitudinal information on various aspects – education, labour, income and wealth, social class, *etc.* –, at the individual and the household level. The cons are small sample size and, for the Survey of the Bank of Italy, the far from satisfactory features of its panel component.

Some studies show interesting attempts to overcome some of the limitations of available data sources. These attempts proceed along two lines. A couple of studies succeeded in integrating different administrative archives (Caruso & Pisauro, 2005; Paggiaro, Rettore & Trivellato, 2009). A few studies were able to design – or resort to – *ad hoc* surveys, tailored to the information requirements of the evaluation exercise (Ichino, Mealli & Nannicini 2005; Battistin & Rettore, 2002; Berliri, Bulgarelli & Pappalardo, 2002).

[13] A further issue, to some extent peculiar to Italy, is the unreasonably restricted legislation on data access for research purposes. See Sestito & Trivellato (2011).

4.3. *Sensitivity analyses and over-identification tests*

I already pointed out the importance of over-identification tests for a credible appraisal of the effects in observational studies. Their role is critical for most LMP evaluations in Italy, given the limitations the analyst faces because of the two previous points: the fact that all evaluations are retrospective – thus they have to rely on fairly stringent identifying assumptions – and frequently face data constraints.

In order to provide some clue on that point, I provide an overall, subjective assessment of the credibility of the identifying restriction and over-identification checks (sensitivity analyses, over-identification tests, *etc.*), expressed on a five-point scale (see Table 1, column "Identification issues" and footnote (a)). If one is willing to rely on my assessment, in several papers the issue has been given just ordinary, in some instances less than satisfactory consideration. Obviously, this casts (sometimes severe) doubts on the credibility of the results.

Attention to problems arising from questionable assumptions is in Ichino, Mealli & Nannicini (2005), though they hardly succeed in finding a convincing solution. The focus on over-identification tests is growing in recent papers: see Martini & Mo Costabella (2007), and Paggiaro, Rettore & Trivellato (2009; 2010).

The increasing concern about the identification issue, sensitivity analyses and over-identification tests shows an appreciable side effect in terms of methodologically oriented contributions to the topic. Ichino, Mealli & Nannicini (2008) come back to the issue of robustness of the selection on observables' assumption in a paper centred on sensitivity analyses of matching estimators, still using the effect of TWA employment in Italy as case-study. Battistin & Rettore (2008) exploit the availability of ineligibles and eligible non participants as a double comparison group in a RDD for specifying meaningful over-identification tests.

4.4. *Possibly heterogeneous effects*

Distribution effects frequently matter. If the returns from programme participation are heterogeneous across units, the overall ATT provides poor, possibly misleading information. Quite simply, treatment effect heterogeneity can be investigated by considering how average treatment effects vary across groups that share the same observable characteristics W, $E\,[Y^T - Y^{NT} \mid D=1, W=w]$, and then tracing the distribution of the effects over the values w. Obviously, this strategy rests on the assumption that the treatment effect heterogeneity of interest occurs with respect to observables. Operationally, it is interesting mainly when the characteristics W are discrete: in that case it boils down to stratification[14].

[14] One may want to go further and get distribution parameters of policy interest: *e.g.*, the proportion of people taking the programme who benefit from it or the distribution of gains, $(Y^T - Y^{NT})$, at selected values of the non-treatment status. To this purpose sensible restrictions should be imposed, in order to identify the joint distribution (Y^T, Y^{NT}) moving from the marginal distributions of the potential outcomes, taken as identified. No one of the reviewed studies takes this route, which rests definitely beyond the scope of this paper; see Battistin & Fort (2008) for a concise, insightful expository note.

It is also data demanding, since adequate sample size by strata is needed[15].

Heterogeneity of the effects is likely to be quite important for LMPs in Italy, because of the large cultural, social and economic differences among groups and areas. Most of the reviewed studies explore heterogeneous effects through stratification by a single, often dichotomous variable, because of sample size constraints.

The geographic area, basically the North & Center *vs.* South divide (or the divide by regions from the two areas), appears to be the most pervasive factor of heterogeneity of the effects. This turns out neatly for TWA employment (Ichino, Mealli & Nannicini, 2005), for temporary contracts (Barbieri & Sestito, 2008; Paggiaro, Rettore & Trivellato, 2010), and for the reform of PES (Loriga & Naticchioni, 2009).

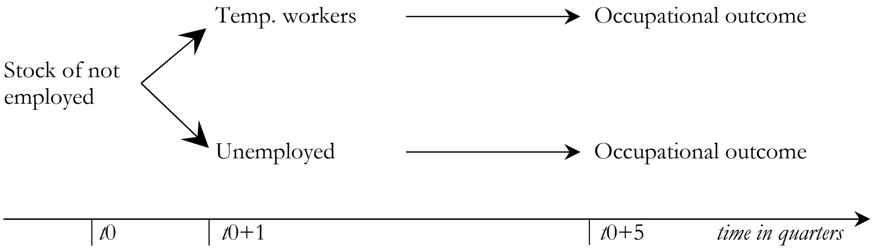
Gender also plays a significant role in some policies and contexts. This is the case again for labour contracts (Barbieri & Sestito, 2008), and for the LiMo programme in the Veneto region – women aged more than 40 significantly postpone re-entry at work – (Paggiaro, Rettore & Trivellato, 2009).

In assessing the effects of temporary contracts *vs.* unemployment on short-term employment outcomes, Barbieri & Sestito (2008) devote peculiar attention to possibly heterogeneous effects. They stratify the ATT estimates separately by geographic area, gender, two age groups, two educational levels, provinces by unemployment rate. Unfortunately, they find no significant differences at all, the reason being inadequate sample size. Indeed, when a much larger sample size is exploited (Paggiaro, Rettore & Trivellato, 2010), one gets significant differences jointly by area and gender

4.5. *Evidence from a case-study on the issues above*

Some additional results from this last research, still in progress, are worth to be presented, because they demonstrate the relevance of the issues just discussed. For their study of the effects of temporary employment regulations, Paggiaro, Rettore & Trivellato (2010) – henceforth PR&T – take the evaluation design from Barbieri & Sestito (2008) – henceforth B&S. That is, they contrast the quarterly flow into temporary employment to the parallel flow into unemployment, and look at occupational outcomes one year ahead, as shown in Graph 1. The design is appropriate for taking advantage of the 2-2-2 rotating panel feature of the Italian LFS (see Table 2).

[15] I will not comment on results about heterogeneous effects documented by studies that use parametric, though flexible, models (Gagliarducci, 2005; Berliri, Bulgarelli & Pappalardo, 2002; Contini *et al.*, 2002).



Graph 1 – Design of the evaluation study.

TABLE 2

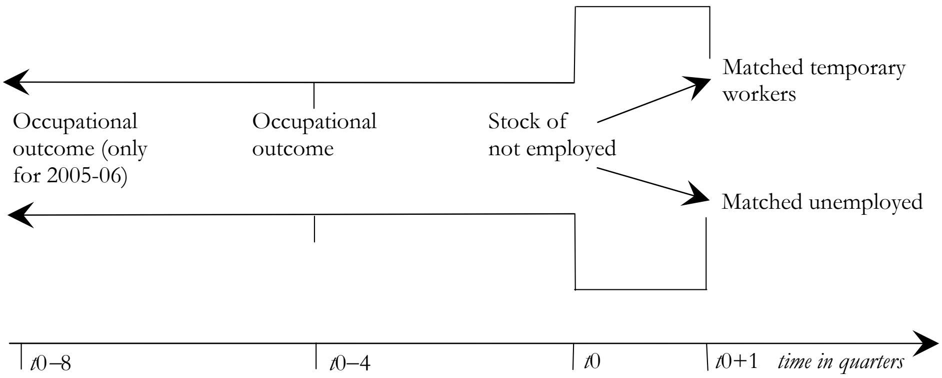
The rotation scheme of the Italian LFS [(a)]

Year	*Quarter*	*Rotation groups*							
2004	4	***A***							
2005	1	A	B						
	2		B	C					
	3			C	D				
	4	***A***			D	**E**			
2006	1	***A***	B			**E**	F		
	2		B	C			F	G	
	3			C	D			G	H
	4				D	E			H
2007	1					**E**	F		
	2						F	G	
	3							G	H
	4								H

[(a)] As an example, in rotation group E the three waves (*t*0, *t*0+1, *t*0+5) used for impact analysis are in bold; in rotation group *A* the three waves used for the backward test are in italics bold.

PR&T improve on B&S in three respects. (i) They get larger samples from the LFS, by pooling series of eight successive short panels. (ii) They carry out the analyses for three two-year periods – 1995-96, 2001-01, 2005-06 –, that differ in two aspects: increasingly open regulations on temporary contracts; a richer information on work experience and the actual job, available just for 2005-06 from the CLFS (Istat, 2004). (iii) Under the evaluation strategy used – matching on the observables –, they offer an over-identification test for the ignorability of treatment status. It exploits the peculiar LFS's sampling scheme outlined in Table 2, and tests the ignorability assumption on an independent sample from the same population used to identify the causal effect by going backward[16]. A representation of the time design of the backward test is in Graph 2.

[16] To exemplify, look at Table 2 and stick to the E sample. The *A* sample represents the very same population as the E one in the last quarter of 2005 (*t*0) and the first quarter of 2006 (*t*0+1). If we apply the same matching strategy to the last two waves of the *A* sample as we do for the first two waves of the E sample, we end up with two couples of treatment and control groups alike up to sampling variability. Then, to evaluate the causal parameter of interest we collect X in the first wave (*t*0), observe the treatment status D in the second wave (*t*0+1), compare the outcomes across the matched groups in the fourth wave (*t*0+5) (see Graph 1); for the backward test we collect X in the third wave (*t*0), observe the treatment status D in the fourth wave (*t*0+1), compare the outcomes across the matched groups in the first wave (*t*0 – 4) (see Graph 2).



[(a)] Source: Paggiaro, Rettore & Trivellato (2010).

Graph 2 – Design of the backward test [(a)].

When used to gauge the periods and groups for which one can draw credible inferences on the causal effect of temporary contracts, the specification test yields a neat answer. The matching estimator systematically fails to pass the test for the first two periods, otherwise stated, when the matching estimator utilizes the relatively poor set of variables provided by the QLFS. By contrast, as the set of matching variables substantially improves, which happens for 2005-06 with data from the CLFS, the estimator survives the test. This evidence demonstrates two previous statements: over-identification tests are vital to validate results from retrospective evaluations; to that purpose, adequate data are a crucial ingredient.

Looking at the results for 2005-06, by pooling eight successive short panels PR&T get a reasonably large sample size, that allows them to produce informative estimates of treatment effect jointly by gender an geographic area – North-Centre *vs.* South – (see Table 3).

TABLE 3

Estimates of the causal effects, 2005-06 sample, by gender and area [(a)] [(b)]

Outcome at t0+5	*Men North-Centre*		*Men South*		*Women North-Centre*		*Women South*	
	ATT	Sign.	ATT	Sign.	ATT	Sign.	ATT	Sign.
Employment rate	26.99	***	27.24	***	29.98	***	33.53	**
Permanent empl. rate	5.57	*	-6.55	**	3.15		3.81	
Satisfactory empl. rate	14.57	***	0.15		9.20	***	7.64	***
No. treated	376		459		568		442	
No. controls	711		1,766		1,168		1,802	

[(b)] Source: adapted from Paggiaro, Rettore & Trivellato (2010).
[(a)] Significance level: *** 1%; ** 5%; *10%.

ATT estimates by strata reveal interesting patterns. Let us focus on the two major outcome variables: "Permanent (≡ open-ended) employment" and "Satisfactory employment" transition rates, respectively. For men the *ATT* estimates are appreciably polarized: the average effect of entering a temporary contract, with respect to unemployment, is significantly positive in the North-Centre

(+ 5,6% for Permanent employment and 14.6% for Satisfactory employment), while it is decidedly negative (– 6,6%) or negligible, respectively, in the South. For women, *ATT* estimates turn out to be even across the two areas and systematically less pronounced; indeed, for them the transition rate to Permanent employment is not significant.

5. SOME SUBSTANTIVE EVIDENCE

Many reasons, among which those given in the previous Section, make it problematic to draw robust substantive evidence about the causal effects of the Italian LM programmes. Indeed, the problem is not confined to the Italian case. Reviews at multi-country[17] and national[18] level do not offer systematically neat indications. This partly depends on the difficulty to ascertain programme effects because of deficiencies of the evaluations studies. For instance, inadequate sample size might be responsible for the inability to detect small effects and, faced with treatment effect heterogeneity, to assess the effects for relevant sub-groups of beneficiaries.

Among the substantive reasons why many LMPs are often ineffective, two are frequently stressed: *pro-capite* expenditures are inadequate to overcome the deficits of participants; several LMPs are poorly targeted (again a feature that results in impact heterogeneity). As for some broad classes of LMPs, it is reasonable to cautiously sum up the available evidence as follows; changes in regulations – aimed at increasing flexibility – have moderate positive effects on employment outcomes, while effects on wages and productivity are ambiguous; subsidized (especially public) jobs are usually ineffective; the impact of vocational and on-the-job training programmes heavily depends on how they are designed, with dominant "lock-in" effects for traditional, long, poorly targeted programmes and moderately positive effects for short, well-targeted interventions; start-up programmes and welfare-to-work policies frequently work, the more they are designed – and consistently implemented – according to the logic of mutual obligations (≡ rights and duties, conditionality) with effective sanctions the better.

These broad conclusions appear to be appropriate also for the last fifteen years of Italian LMPs. They can be usefully complemented by some further comments, largely specific to the Italian case.

As already pointed out, the bulk of interventions were changes in labour market regulations – confined to new entrants and targeted to flexibility –, while the welfare system was not modified correspondingly. Overall, the effect of more flexible labour contracts was a moderate growth of (mainly temporary) employment. Evidence from studies on the aggregate dynamics of labour flows and

[17] See Heckman, Lalonde & Smith (1999, pp. 2043-2080), Martin & Grubb (2001), Kluve *et al.* (2007), Bassanini, Nunziata & Venn (2008), Card, Kluve & Weber (2009), and Martin & Scarpetta (2011), among others.

[18] At the national level, an interesting case are "Hartz reforms" in Germany, which have been systematically evaluated. For concise reviews see Caliendo (2009) and Martini and Trivellato (2011, pp. 82-90 and 160-163).

wages shows that for new entrants – largely with temporary contracts – there was an increase of labour mobility and a decline of wages (the initial wage and the age-wage profile), and hints to negative effects on productivity (Giorgi *et al.*, 2001).

The only (supposedly) appreciable institutional change was the reform of PES. In fact, its effectiveness was scanty, reasonably because it was mainly a "law on the books", not a reform "in action".

The main LMPs introduced – or modified – consist of a mix of active and passive measures. There effects are generally meagre. This partly depends on the poor – possibly inconsistent – policy design (for instance, this is the case for the provision of LiMo for workers aged 50 years or more). Besides, the implementation of these programmes is flawed, for various reasons. (i) Usually the programme is designed at the national level and put into operation at local levels, with varying protocols, by agencies/agents that might have different goals. (ii) As a rule, active and passive measures of a policy are administered by different agencies (typically, regional or local agencies deliver the active measures, while the passive measures are managed by INPS), with lack of coordination. (iii) Frequently active measures are weak – sometimes they are just "on the books" –, because of the poor functioning of PES. (iv) Then, it come with no surprise that sanctions are rarely enforced.

Department of Statistics, University of Padova
and IRVAPP

UGO TRIVELLATO

REFERENCES

B. ANASTASIA, M. MANCINI E U. TRIVELLATO (2009), *Il sostegno al reddito dei disoccupati: note sullo stato dell'arte. Tra riformismo strisciante, inerzie dell'impianto categoriale e incerti orizzonti di flexicurity*, Roma, ISAE Working Paper n. 112.

B. ANASTASIA, A. PAGGIARO E U. TRIVELLATO (2011), *Gli effetti delle riforme nella regolazione e nel welfare del lavoro sulle disuguaglianze generazionali*, in A. Schizzerotto, U. Trivellato e N. Sartor (Eds.), *Generazioni disuguali. Le condizioni di vita dei giovani di ieri e di oggi: un confronto*, Bologna, il Mulino [forthcoming].

G. BARBIERI, P. GENNARI, G. LINFANTE, E. RUSTICHELLI AND P. SESTITO (2003), *Valutare i servizi pubblici per l'impiego: implicazioni delle riforme, attivismo dei servizi e chances lavorative degli utenti*, "Politica Economica", 19 (3), pp. 343-372.

G. BARBIERI, P. SESTITO (2008), *Temporary workers in Italy: Who are they and where they end up*, "Labour", 22 (1), pp. 127-166.

P. BARBIERI, S. SCHERER (2007), "Vite svendute. Uno sguardo analitico sulla costruzione sociale delle prossime generazioni di esclusi", *Polis* 21 (3), pp. 431-459.

F. BARCA (2009), *An agenda for a reformed cohesion policy. A place-based approach to meeting European Union challenges and expectations*, Brussels, Independent Report prepared at the request of Danuta Hübner, Commissioner for Regional Policy, April 2009 [http:// ec.europa.eu/regional_policy/policy/future/ barca_en.htm].

A. BASSANINI, L. NUNZIATA, D. VENN (2008). *Job protection legislation and productivity growth in OECD countries*, Bonn, IZA Discussion Paper No. 3555.

E. BATTISTIN, M. FORT (2008), *What's missing from policy evaluation: Identification and estimation of the distribution of treatment effects*, in Società Italiana di Statistica, "Atti della XLIV Riunione Scientifica", Padova, Cleup, pp. 127-134.

E. BATTISTIN, E. RETTORE (2002), *Testing for programme effects in a regression discontinuity design with imperfect compliance*, "Journal of the Royal Statistical Society, A", 165 (1), pp. 39-57.

E. BATTISTIN, E. RETTORE (2008), *Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs*, "Journal of Econometrics", 142 (2), pp. 715-730.

R. BERK (2005), *Randomized experiments as the bronze standard*, Paper 2005080201, Department of Statistics, UCLA [http://repositories.cdlib. org/uclastat/papers/2005080201].

C. BERLIRI, A. BULGARELLI, C. PAPPALARDO (2002), *Valutazione della qualità della formazione professionale attraverso la stima dell'occupabilità*, "Lavoro e Relazioni Industriali", 1, pp. 25-44.

F. BERTON, F. DEVICIENTI, L. PACELLI (2008), *Are temporary jobs a port of entry into permanent employment? Evidence from matched employer-employee data*, Moncalieri, LABORatorio Riccardo Revelli - Center for Employment Studies Working Paper No. 79.

I. BISON, E. RETTORE, A. SCHIZZEROTTO (2010), "La riforma Treu e la mobilità contrattuale in Italia. Un confronto tra coorti", in D. Checchi (ed.), *Immobilità diffusa. Perché la mobilità intergenerazionale è così bassa in Italia*, Bologna, il Mulino, pp. 267-296.

R. BLUNDELL, M. COSTA DIAS (2009), *Alternative approaches to evaluation in empirical microeconomics*, "Journal of Human Resources", 44 (3), pp. 565-640.

M. CALIENDO (2009), *Income support systems, labor market policies and labor supply. The German experience*, Bonn. IZA Discussion Paper No. 4665.

D. CARD, J. KLUVE, A. WEBER (2009), *Active labor market policy evaluations: A meta-analysis*, Bonn, IZA Discussion Paper No. 4002.

E. CARUSO, G. PISAURO G. (2005), *Licenziamenti definitivi o temporanei? Durata della disoccupazione nelle Liste di mobilità tra nuovi e vecchi datori di lavoro*, "Politica Economica", 21 (3), pp. 361-399.

COMMISSION OF THE EUROPEAN COMMUNITIES (2007), *Towards common principles of flexicurity: More and better jobs through flexibility and security*, SEC (2007)861, Brussels.

B. CONTINI, F. CORNAGLIA, C. MALPEDE, E. RETTORE (2002), *Measuring the impact of the Italian CFL programme on the job opportunities for the youths*, in O. Castellino and E. Fornero (Eds.), *Pension policy in an integrating Europe*, Cheltenham, Edward Elgar, pp. 85-105.

R. ERCOLI, A. GUELFI (2008), *Schede sinottiche di studi di valutazione degli effetti di sussidi alle imprese finalizzati all'incremento dell'occupazione*, in U. Trivellato (ed.), *Analisi e proposte in tema di valutazione degli effetti di politiche del lavoro*, Commissione di Indagine sul Lavoro, Rapporto n. 10, Roma, Cnel, pp. 131-169 [http://www.portalecnel.it/Portale/IndLavrapporti Finali.nsf/vwCapitoli? OpenView&Count=40].

S. GAGLIARDUCCI (2005), *The dynamics of repeated temporary jobs*, "Labour Economics", 12 (4): 429-448.

P. GARIBALDI, L. PACELLI, A. BORGARELLO (2004), *Employment protection legislation and the size of firms*, "Giornale degli Economisti e Annali di Economia", 63 (1), pp. 33-68.

F. GIORGI, A. ROSOLIA, R. TORRINI, U. TRIVELLATO (2011), *Mutamenti tra generazioni nelle condizioni lavorative giovanili,* in A. Schizzerotto, U. Trivellato e N. Sartor (Eds.), *Generazioni disuguali. Le condizioni di vita dei giovani di ieri e di oggi: un confronto*, Bologna, il Mulino [forthcoming].

J.J. HECKMAN, V.J. HOTZ (1989), *Choosing among alternative non-experimental methods for estimating the impact of social programmes: the case of manpower training*, "Journal of the American Statistical Association", 84 (408), pp. 862-874.

J.J. HECKMAN, R.J. LALONDE, J.A. SMITH (1999), *The economics and econometrics of active labor market programs*, in O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics – Vol. 3A*, Amsterdam, Elsevier, pp. 1865-2097.

A. ICHINO, F. MEALLI, T. NANNICINI (2005), *Temporary Work Agencies in Italy: A springboard to permanent employment?*, "Giornale degli Economisti e Annali di Economia", 64 (1), pp. 1-27.

A. ICHINO, F. MEALLI, T. NANNICINI (2008), *From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity?*, "Journal of Applied Econometrics", 23 (3), pp. 305-327.

G.W. IMBENS, T. LEMIEUX (2008), *Regression Discontinuity Designs: A guide to practice*, "Journal of Econometrics" 142 (2), pp. 615-635.

G.W. IMBENS, J.M. WOOLDRIDGE (2009), *Recent developments in the econometrics of program evaluation*, "Journal of Economic Literature", 47 (1), pp. 5-86.

ISTAT (2004), *La nuova rilevazione sulle forze di lavoro - Contenuti, metodologie, organizzazione*, Roma.

J. KLUVE, ET AL. (2007), *Active labor market policies in Europe: Performance and perspectives*, Berlin-Heidelberg, Springer.

D.S: LEE (2008), *Randomized experiments from non-random selection in U.S. House elections*, "Journal of Econometrics", 142 (2), pp. 675-697.

J.P. MARTIN, D. GRUBB (2001), *What works and for whom: A review of OECD countries' experience with Active Labour Market Policies*, "Swedish Economic Policy Review", 8 (1), pp. 9-56.

J.P. MARTIN, S. SCARPETTA (2011), *Setting it right: Employment protection, labour reallocation and productivity*, Bonn, IZA Policy Paper No. 27.

A. MARTINI (2009), *RCT to test a program to place mentally ill patients into permanent jobs*, Torino, Progetto Valutazione [mimeo; restricted].

A. MARTINI, L. MO COSTABELLA (2007), *Una valutazione degli effetti indesiderati dell'istituto della mobilità su imprese e lavoratori*, "Politica Economica", 23 (3), pp. 259-288.

A. MARTINI, U. TRIVELLATO (2011), *Sono soldi ben spesi? Perché e come valutare l'efficacia delle politiche pubbliche*, Venezia, Marsilio.

J. MCCRARY (2008), *Manipulation of the running variable in the Regression Discontinuity Design: A density test*, "Journal of Econometrics", 142 (2), pp. 698-714.

P. NATICCHIONI, S. LORIGA (2008), *Short and long term evaluations of Public Employment Services in Italy*, Discussion Paper 2008-30, Département de Sciences Économiques, Université Catholique de Louvain.

OECD (1994), *The OECD jobs study: Facts, analyses, & strategies*, Paris.

OECD (various years), *Employment Outlook*, Paris.

A. PAGGIARO, E. RETTORE, U. TRIVELLATO (2008), *The effect of a longer eligibility to a labour market programme for dismissed workers*, "Labour", 23 (1), pp. 37-66.

A. PAGGIARO, E. RETTORE, U. TRIVELLATO (2010), *The effect of experiencing a spell of temporary employment vs. a spell of unemployment on short-term labour market outcomes*, paper presented at the 10th Econometric Society World Congress, Shanghai, August 2010 [a preliminary version appeared as IRVAPP Progress Report No. 2009-03, Trento].

S. PIRRONE, P. SESTITO (2006), *Disoccupati in Italia. Tra Stato, Regioni e cacciatori di teste*, Bologna, il Mulino.

S. PIRRONE, P. SESTITO (2009), *Gli indirizzi della regolazione e delle politiche del lavoro: ricostruzione storica e questioni aperte*, in U. Trivellato (ed.), *Regolazione, welfare e politiche attive del lavoro*, Commissione di Indagine sul Lavoro, Rapporto n. 11, Roma, Cnel, pp. 109-164 [http:// www.portalecnel.it/Portale/IndLavrapportiFinali.nsf/vwCapitoli?OpenView&Count=40].

P.R. ROSEMBAUM (1984), *From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment*, "Journal of the American Statistical Association", 79 (385), pp. 41-48.

P.R. ROSEMBAUM (1987), *The role of a second control group in an observational study*, “Statistical Science”, 2 (3), pp. 292-306.

F. SCHIVARDI, R. TURRINI (2008), *Identifying the effects of firing restrictions through size-contingent differences in regulation*, “Labour Economics”, 15 (3), pp. 482-511.

P. SESTITO (2002), *Il mercato del lavoro in Italia. Com’è, come sta cambiando*, Roma-Bari, Laterza.

P. SESTITO, P. GENNARI, G. BARBIERI, G. LINFANTE, E. RUSTICHELLI (2003), *Valutare i servizi pubblici per l’impegno: implementazione della riforma, attivismo dei servizi e chances lavorative degli utenti*, “Politica Economica”, (3), pp. 343-372.

P. SESTITO, U. TRIVELLATO (2011), *Indagini dirette e fonti amministrative: dall’alternativa all’ancora incompiuta integrazione*, “Rivista di Politica Economica”, July-September 2010-11 [forthcoming].

U. TRIVELLATO, S. ZEC (2008), *Schede sinottiche di studi di valutazione degli effetti di politiche del lavoro*, in U. Trivellato (ed.), *Analisi e proposte in tema di valutazione degli effetti di politiche del lavoro*, Commissione di Indagine sul Lavoro, Rapporto n. 10, Roma, Cnel, pp. 57-130. [http://www.portalecnel.it/Portale/IndLavrapportiFinali.nsf/vwCapitoli?OpenView&Count=40]

SUMMARY

Fifteen years of labour market regulations and policies in Italy: What have we learned from their evaluation?

During the last fifteen years a notable number of labour market interventions took place in Italy, under two main headings: new regulations and implementation of active and/or passive programmes. The paper reviews a fair number of subjectively selected impact evaluation studies published in the last ten years. The framework is provided by counterfactual analysis; preliminary, its barebones are sum up. The focus of the review in on two aspects. First, a blend of empirical and analytical issues critical for credible impact evaluations are discussed. They emerge from the fact that the reviewed studies are carried out in an observational setting, and refer to prospective *vs.* retrospective evaluation, availability (or lack) of adequate data, over-identification tests in order to corroborate (or falsify) the identifying restriction on which the evaluation strategy rests, and heterogeneous effects. Second, the substantive evidence offered by the reviewed studies is summarized. It points to generally minor policy effects; some tentative explanations for that lack of effectiveness are suggested.