A Note on the Relation of Weighting and Matching Estimators

Michael Lechner

September 2007 Discussion Paper no. 2007-34

Department of Economics University of St. Gallen

Editor:	Prof. Jörg Baumberger University of St. Gallen Department of Economics Bodanstr. 1 CH-9000 St. Gallen Phone +41 71 224 22 41 Fax +41 71 224 28 85 Email joerg.baumberger@unisg.ch
Publisher:	Department of Economics University of St. Gallen Bodanstrasse 8 CH-9000 St. Gallen Phone +41 71 224 23 25 Fax +41 71 224 22 98
Electronic Publication:	http://www.vwa.unisg.ch

A Note on the Relation of Weighting and Matching Estimators

Michael Lechner

Author's address: Professor Dr. Michael Lechner
SIAW-HSG
Bodanstrasse 8
9000 St. Gallen
Tel. +41 71 224 2350
Fax +41 71 224 2298
Email michael.lechner@unisg.ch
Website www.siaw.unisg.ch/lechner

Abstract

This paper compares the inverse-probability-of-selection-weighting estimation principle with the matching principle and derives conditions for weighting and matching to identify the same and the true distribution, respectively. This comparison improves the understanding of the relation of these estimation principles and allows constructing new estimators.

Keywords

Matching, inverse-of-selection-probability weighting, treatment evaluation, unconfoundedness

JEL Classification

C21, C13, C14.

1 Introduction*

This paper considers the estimation problem that occurs when the distribution of a random variable Y in population t cannot be directly observed but must be learned from its distribution in population o. Such a situation is typical in the so-called treatment evaluation literature: For example in an evaluation of the effects of a public training programme for unemployed, we can learn the post-programme labour market outcome (Y) of programme participants (population t) directly from the data (see the survey Heckman, LaLonde, and Smith, 1999). However, the outcome they would have realized had they not participated in the programme is not observable and has to be learned from 'comparable' non-participants (population o). For this purpose, the population of nonparticipants has to be adjusted (reweighted) to become 'comparable' to the participants in some dimensions. For strategies that are based on adjustments leading to similar distributions of observables for populations t and o, matching methods as well as inverse-of-selection-probability weighting schemes (IPW) are very popular. Imbens (2004) provides an excellent review of many aspects and variants of these classes of estimators.

In this paper, we compare the IPW estimation principle with the matching principle. Conditions on the properties of the respective weights are derived such that weighting and matching on those weights identify the same distribution. Further conditions are provided such that the distribution identified is indeed the distribution of interest. The objective of this paper is not to analyse the small sample (e.g. Frölich, 2004) or large sample (e.g. Abadie and Imbens, 2006, Firpo, 2007) properties of the different estimators that are applicable in such a context, but to investigate the conditions the weights have to fulfil in the population (or in an indefinitely large i.i.d. sample).

* I am also affiliated with ZEW, Mannheim, CEPR and PSI, London, IZA, Bonn, and IAB, Nuremberg.

Under suitable regularity conditions, estimators using consistent estimators of such weights appropriately will usually be consistent as well.

This comparison does not only deepen the understanding of the relation of those two important and frequently used estimation principles, but does also allow constructing new consistent estimators. For example, if a weighting estimator is known to be consistent and its weights fulfil the conditions derived in this paper, then a consistent matching-type estimator can be constructed solely based on such weights.

A comparison of these estimation principles from this population perspective has not yet received much attention in the literature. There are some estimator specific asymptotic and small sample results of estimators based on IPW and stratification, which is very similar to matching, by Lunceford and Davidian (2004). However, their comparison is for a case when both estimation principles necessarily identify the same density. Furthermore, there are formal comparisons (e.g. Ichimura and Taber, 2001, Frölich and Lechner, 2004), as well as more informal comparisons (e.g. Hogan and Lancaster, 2004), of instrumental variable methods with weighting and matching approaches. However, they use a different perspective and the method of instrumental variables is not the focus of this paper.

In the next section same notation is introduced, the objects of interest are defined, and regularity conditions are imposed to simplify the subsequent analysis. Section 3 defines the estimation principles and Sector 4 compares the matching and weighting estimators. Section 5 concludes.

2 Notation, targets for the estimation and sampling frame

There are two subpopulations depending on the value of the random variable *S*, namely the target population (*S*=*t*) and the observed population (*S*=*o*).[1] Realisations of *Y,* which is the random variable of interest, are only observed in the target population, whereas realisations of a set of confounding variables *X* and weights *W* are observed in both subpopulations. We are interested in learning the distribution of *Y* in the subpopulation in which it is not observed, i.e. $F_{Y|S}(y,t)$. To do so, the information in *X* and *W* will be combined with the corresponding distributions of *Y* in the population in which realisation of *Y* are observable.

In Assumption 1, the sampling scheme is formalized and enough regularity and support is assumed to concentrate on the key issues of comparing the two estimation principles.

Assumption 1 (sampling frame, common support, regularity)

(a) Sampling: There is a random sample of size *N* ($\{s_i, x_i, y_i\}_{i=1}^{N}$) from the joint distribution of the random variables (*S, X, Y*). In the subsample with $s_i = t$, the values of y_i are not observable.

(b) Support and regularity: Assume that all densities and moments that are of interest are finite and nonzero.

Two remarks are in order concerning Assumption 1: First, the i.i.d. sampling assumption is not important for what follows. What is important is that y_i is observed in one subsample, but not in the other. Second, part b) is obviously overly restrictive, but it allows concentrating on the main issues in this paper without additional notation.

[1] As a convention, capital letters denote random variables and small letters denote either realisations or specific values of those random variables. Furthermore, $F_{A|B}(a,b)$ denotes the cumulative distribution functions of *A*

Having defined the general model, which is very simple, the following section presents the different population problems that matching and direct weighting estimators solve.

3 Principles of direct weighting and a matching estimation

It is the idea of the weighting estimator to take the empirical mean in the observable population over an individual specific weight times a function of the observed random variables of interest in the subpopulation in which y is observed. This idea is formalized in Definition 1.

Definition 1 (estimation principle of weighting estimator)

The estimation principle of the weighting estimator is defined for a one dimensional random variable *W such that:* $F^{W}_{Y|S}(y,t) = \underset{W|S=o}{E}\left[wF_{Y|W,S}(y,w,o) \right]$.

At this stage of generality, $F^{W}_{Y|S}(y,t)$ may or may not be equal to the true distribution $F_{Y|S}(y,t)$ and w may or may not correspond to some function of selection probabilities. The idea of using inverse selection probability weighting is probably due to Horowitz and Thomson (1952), but has since then been analysed by many others.

A second set of weights define a matching-type estimator, which, like all estimators, has an interpretation as a weighting estimator as well. The difference between direct weighting and matching is how the weights are defined. For direct weighting, the weights are directly defined such that their expectation over the observable population fulfils a certain restriction. For matching, the weights are implicitly determined from the distribution of some variables in the population for which Y is not observable. We call those variables X.

---

conditional on B evaluated at the value a for B = b. $f_{A|B}(a,b)$ is the corresponding density. If A discrete, $f_{A|B}(a,b)$ is the corresponding mass function.

Definition 2a (matching principle based on covariates)

The vector of random variables X defines the principle of matching estimating such that:

$$F_{Y|S}^{M(x)}(y,t) = \underset{X|S=t}{E}\left[F_{Y|X,S}(y,x,o)\right].$$

As for weighting, at this level of generality there is no need to assume that $F_{Y|S}^{M(x)}(y,t)$ = $F_{Y|S}(y,t)$. The idea of adjusting observed (confounding) variables directly can at least be traced back to Fechner (1860), and is discussed in Wilks (1932) and Rubin (1979). The conditions required for $F_{Y|S}^{M(x)}(y,t)$ = $F_{Y|S}(y,t)$ are intensively discussed in the literature as well. They typically come under the heading of the conditional independence or no confounding assumption (e.g., see Cochrane and Chambers, 1965, Rubin, 1974). Estimation methods for matching estimation as well as weighting under no confounding are extensively discussed by Imbens (2004).

Analogously to the weighting principle, Definition 2b) defines a matching estimator that is based on a one-dimensional covariate, which we call W as before.

Definition 2b (matching principle based on weights)

The random variable W (weight) defines the principle of weight-based matching estimating such that: $F_{Y|S}^{M(w)}(y,t) = \underset{W|S=t}{E}\left[F_{Y|W,S}(y,w,o)\right]$.

The key distinction between these three definitions is the different dimension of the random variables W and X and how they are used to adjust the distribution. A particular example in which a matching estimator is consistent for a one-dimensional conditioning variable (the so-called propensity score, defined as $P(S=t \mid X=x, S \in \{t,o\})$) has been discussed by Rosenbaum and Rubin (1983).

The following analysis sheds more light on the relation between these estimation principles.

4 The relation of the estimation principles

The first question we analyse in this section is whether there are weights for which both estimation principles identify the same distribution. Theorem 1 establishes that such weights exist (under suitable support and regularity conditions) and that they have a well-known form.

Theorem 1 (weights that lead to equivalence of weighting and matching)

a) If Assumption 1 holds, then the following weights lead to $F^{W}_{Y|S}(y,t)$ = $F^{M(x)}_{Y|S}(y,t)$:

$$w = w(x) = \frac{P(S=t \mid X=x)}{P(S=o \mid X=x)}\frac{P(S=o)}{P(S=t)} = \frac{f_{X|S}(x,t)}{f_{X|S}(x,o)}.$$

b) If Assumption 1 holds, then the following weights lead to $F^{W}_{Y|S}(y,t)$ = $F^{M(w)}_{Y|S}(y,t)$:

$$w = \frac{P(S=t \mid W=w)}{P(S=o \mid W=w)}\frac{P(S=o)}{P(S=t)} = \frac{f_{W|S}(w,t)}{f_{W|S}(w,o)}.$$

The proof of this theorem is relegated to Appendix A.1.

Note that this theorem does not establish that either a matching or a weighting estimator based on these weights identifies the true distribution of interest. However, if one of the estimation principles leads to the true distribution using such weights, then Theorem 1 implies that the other estimation principle recovers the true distribution as well.

The weights appearing in Theorem 1a) and 1b) are so-called inverse probability of selection (IPW) weights either as function of *X* or of the summary measure *W*. One such summary measure that fulfils this criterion in the binary treatment model under the unconfoundedness (Rubin, 1974) assumption is the already mentioned propensity score (see Rosenbaum and Rubin, 1983).

Next, Theorem 2 provides conditions under which the matching and weighting principles identify the true distribution.

Theorem 2 (weights leading to consistent matching estimation)

a) If the vector *X* fulfils the following condition, then the matching principle identifies $F_{Y|S}(y,t)$:

$$\underset{X|S=t}{E}\left[F_{Y|X,S}(y,x,t)-F_{Y|X,S}(y,x,o)\right]\overset{!}{=}0\text{, or}$$

$$\underset{X|S=o}{E}\left\{\frac{P(S=t\mid X=x)}{P(S=o\mid X=x)}\frac{P(S=o)}{P(S=t)}\left[F_{Y|X,S}(y,x,t)-F_{Y|X,S}(y,x,o)\right]\right\}\overset{!}{=}0\,.$$

b) If the weights *W* fulfil the following condition, then the matching principle identifies $F_{Y|S}(y,t)$:

$$\underset{W|S=t}{E}\left[F_{Y|W,S}(y,w,t)-F_{Y|W,S}(y,w,o)\right]\overset{!}{=}0\text{, or}$$

$$\underset{W|S=o}{E}\left\{\frac{P(S=t\mid W=w)P(S=o)}{P(S=o\mid W=w)P(S=t)}\left[F_{Y|W,S}(y,w,t)-F_{Y|W,S}(y,w,o)\right]\right\}\overset{!}{=}0\,.$$

Noting that $f_{W|S}(w,t) = \frac{P(S=t\mid W=w)}{P(S=o\mid W=w)}\frac{P(S=o)}{P(S=t)}f_{W|S}(w,o)$ and $f_{X|S}(x,t) = \frac{P(S=t\mid X=x)}{P(S=o\mid X=x)}\frac{P(S=o)}{P(S=t)}f_{X|S}(x,o)$,[2] the proof is direct. It is therefore omitted for the sake of brevity. Note that these conditions are less restrictive than directly requiring $F_{Y|X,S}(y,x,t) = F_{Y|X,S}(y,x,o)$ or $F_{Y|W,S}(y,w,t) = F_{Y|W,S}(y,w,o)$. However, any difference between $F_{Y|X,S}(y,x,t)$ and

[2] This result follows directly from $\frac{f_{A,B}(a,b)}{f_{A|B}(a,b)}=\frac{f_{B|A}(b,a)f_A(a)}{f_{A|B}(a,b)}=f_B(b)=\frac{f_{A,B}(a',b)}{f_{A'|B}(a',b)}=\frac{f_{B|A}(b,a')f_A(a')}{f_{A|B}(a',b)}$.

$F_{Y|X,S}(y,x,o)$, or between $F_{Y|W,S}(y,w,t)$ and $F_{Y|W,S}(y,w,o)$, has to be averaged away in the respective distribution of X or W.

For further comparisons with the weighting principle to be discussed below it is interesting to reformulate $F_{Y|W,S}(y,w,t)$ in terms of $F_{Y|W,S}(y,w,o)$, i.e. $F_{Y|W,S}(y,w,t) = \dfrac{P(S=t\,|\,Y=y,W=w)}{P(S=o\,|\,Y=y,W=w)}$ $\dfrac{P(S=o\,|\,W=w)}{P(S=t\,|\,W=w)}$ $F_{Y|W,S}(y,w,o)$. From this property, we restate the conditions of Theorem 2b):

Theorem 2b'

If the weights W fulfil the following condition, then the matching principle identifies $F_{Y|S}(y,t)$:

$$\underset{W|S=o}{E}\left\{\frac{P(S=o)}{P(S=t)}\left[\frac{P(S=t\,|\,Y=y,W=w)}{P(S=o\,|\,Y=y,W=w)}-\frac{P(S=t\,|\,W=w)}{P(S=o\,|\,W=w)}\right]F_{Y|W,S}(y,w,o)\right\}=$$
$$=\underset{W|S=t}{E}\left\{\left[\frac{P(S=o\,|\,W=w)}{P(S=t\,|\,W=w)}\frac{P(S=t\,|\,Y=y,W=w)}{P(S=o\,|\,Y=y,W=w)}-1\right]F_{Y|W,S}(y,w,o)\right\}\stackrel{!}{=}0.$$

If the classical matching assumption is fulfilled conditional on the weights (i.e. $P(S=t\,|\,Y=y,W=w) = P(S=t\,|\,W=w)$ and $P(S=o\,|\,Y=y,W=w) = P(S=o\,|\,W=w)$), or if $\dfrac{P(S=o\,|\,W=w)}{P(S=t\,|\,W=w)} = \dfrac{P(S=o\,|\,Y=y,W=w)}{P(S=t\,|\,Y=y,W=w)}$, then the true distribution of interest is identified. Note that if the conditions of Theorem 2b) or 2b') are satisfied and if the weights have the form as in Theorem 1b), then the weighting principle based on such weights identifies the true distribution of interest as well.

Next, in Theorem 3 we consider the condition such that weighting identifies the true distribution.

Theorem 3 (weights leading to consistent weighting estimation)

If the weights W fulfil the following condition, then the weighting principle identifies $F_{Y|S}(y,t)$:

$$\underset{W|S=o}{E}\left\{\left[\frac{P(S=t\,|\,Y=y,W=w)}{P(S=o\,|\,Y=y,W=w)}\frac{P(S=o)}{P(S=t)}-w\right]F_{Y|W,S}(y,w,o)\right\}=$$

$$=\underset{W|S=t}{E}\left\{\frac{P(S=o\,|\,W=w)}{P(S=t\,|\,W=w)}\left[\frac{P(S=t\,|\,Y=y,W=w)}{P(S=o\,|\,Y=y,W=w)}-w\frac{P(S=t)}{P(S=o)}\right]F_{Y|W,S}(y,w,o)\right\}\overset{!}{=}0.$$

The proof of Theorem 3 is contained in Appendix A.2. Clearly, if $w = \frac{P(S=t\,|\,Y=y,W=w)}{P(S=o\,|\,Y=y,W=w)}\frac{P(S=o)}{P(S=t)}$, this criterion is fulfilled. However, since $P(S=t\,|\,Y=y,W=w)$ cannot be directly learned from the data,[3] more assumptions are needed. They could be of the different types extensively discussed in the literature (e.g. Heckman, LaLonde, and Smith, 1999). For example, if $\frac{P(S=t\,|\,Y=y,W=w)}{P(S=o\,|\,Y=y,W=w)}$ is equal to $\frac{P(S=t\,|\,W=w)}{P(S=o\,|\,W=w)}$, which follows from the matching assumption, then the weights fulfil the conditions for the equality of matching and weighting and matching leads to consistent estimation as well. Note again that the matching and weighting conditions are identical for IPW-selection on observables weights: $w = \frac{P(S=t\,|\,W=w)}{P(S=o\,|\,W=w)}\frac{P(S=o)}{P(S=t)}$.

5 Conclusion

We compare the inverse-probability-of-selection-weighting principle with the matching principle. Conditions on the properties of the respective weights are derived such that weighting and matching on those weights identify the same distribution. Further conditions are provided such that this distribution is the true one. Under suitable regularity conditions, estimators using consistent estimators of such weights appropriately will usually be consistent.

[3] Obviously if $P(S=t\,|\,Y=y,W=w)$ cannot be consistently estimated, $P(S=o\,|\,Y=y,W=w)$ cannot be learned from the data either.

This comparison deepens the understanding of the relation of these estimation principles and allows constructing new consistent estimators. For example, if a weighting estimator is consistent and its weights fulfil the conditions derived in this paper, then a consistent matching-type estimator can be constructed based solely on such weights.

References

Abadie, A., and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects", *Econometrica*, 74, 235-267.

Cochran, W. J., and S. P. Chambers (1965): "The Planning of Observational Studies of Human Populations", *Journal of the Royal Statistical Society, Series* A, 128 (2), 234-266.

Fechner, G. T. (1860), *Elemente der Psychophysik*, Leipzig: Breitkopf und Härtel.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects", *Econometrica*, 75, 259-276.

Frölich, M. (2004): "Finite sample properties of propensity-score matching and weighting estimators," *Review of Economics and Statistics*, 86, 77-90.

Frölich, M., and M. Lechner (2006): "Exploiting regional treatment intensity for the evaluation of labour market policies", 5728, C.E.P.R. Discussion Papers.

Heckman, J. J., R. LaLonde, and J. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs", in: O. Ashenfelter and D. Card (eds.), *Handbook of Labour Economics*, Vol. 3, 1865-2097, Amsterdam: North-Holland.

Hogan, J., and T. Lancaster (2004): "Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies", *Statistical Methods in Medical Research*, 13, 17-48.

Horvitz, J. L., and D. Thomson (1952): "A Generalisation of Sampling without Replacement from a Finite Population", *Journal of the American Statistical Association*, 47, 663-685.

Ichimura, H., and C. Taber (2001): "Propensity-Score Matching with Instrumental Variables", *American Economic Review - Papers and Proceedings*, 91, 2, 119-124.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: a Review", *Review of Economic and Statistics*, 86, 4-29.

Lunceford, J. K., and M. Davidian (2004): "Stratification and weighting via the propensity score in estimation causal treatment effects: a comparative study", *Statistics in Medicine*, 23, 2937-2960.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrica*, 70, 41-50.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66, 688-701.

Rubin, D. B. (1979): "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies", *Journal of the American Statistical Association*, 74, 318-328.

Wilks, S. S. (1932): "On the distribution of statistics in samples from a normal population of two variables with matched sampling of one variable," *Metron,* 9, 87-126.

Appendix A: Proof of Lemmas and Theorems

A.1 Proof of Theorem 1

Start the proof of Theorem 1a) by rewriting both estimation principles using iterated expectations (*I.E.*) to make them more easily comparable:

$$F_{Y|S}^{M(x)}(y,t) = \underset{X|S=t}{E}\left[F_{Y|X,S}(y,x,o)\right] \stackrel{I.E.}{=} \underset{X|S=t}{E}\left[\underset{W|X,S=o}{E} F_{Y|W,X,S}(y,w,x,o)\right].$$

Applying Bayes' Law to the densities of *X* conditional on *S*, we get the following equality:

$$f_{X|S}(x,t) = \frac{P(S=t \mid X=x)P(S=o)}{P(S=o \mid X=x)P(S=t)} f_{X|S}(x,o).$$

Therefore, we can change expectation in the first part of the previous equation to ease comparison between the estimation principles:

$$\underset{X|S=t}{E}\left[\underset{W|X,S=o}{E} F_{Y|W,X,S}(y,w,x,o)\right] \overset{Bayes\ Law}{=} \underset{X|S=o}{E}\left[\frac{P(S=t\mid X=x)P(S=o)}{P(S=o\mid X=x)P(S=t)} \underset{W|X,S=o}{E} F_{Y|W,X,S}(y,w,x,o)\right]=$$

$$= \underset{X|S=o}{E}\left[\underset{W|X,S=o}{E} \frac{P(S=t\mid X=x)P(S=o)}{P(S=o\mid X=x)P(S=t)} F_{Y|W,X,S}(y,w,x,o)\right]$$

$$= \underset{W,X|S=o}{E}\left[\frac{P(S=t\mid X=x)P(S=o)}{P(S=o\mid X=x)P(S=t)} F_{Y|W,X,S}(y,w,x,o)\right].$$

Next, iterated expectations are applied to the weighting principle:

$$F^{W}_{Y|S}(y,t) = \underset{W|S=o}{E}\left[wF_{Y|W,S}(y,w,o)\right] \overset{I.E.}{=} \underset{W|S=o}{E}\left[\underset{X|W=w,S=o}{E} wF_{Y|W,X,S}(y,w,x,o)\right]=$$

$$= \underset{W,X|S=o}{E}\left[wF_{Y|W,X,S}(y,w,x,o)\right].$$

Therefore, the difference between the distributions identified by the two estimation principles is:

$$F^{M(x)}_{Y|S}(y,t) - F^{W}_{Y|S}(y,t) = \underset{W,X|S=o}{E}\left\{\left[\frac{P(S=t\mid X=x)P(S=o)}{P(S=o\mid X=x)P(S=t)} - w\right] F_{Y|W,X,S}(y,w,x,o)\right\}.$$

Therefore, if $w = w(x) = \dfrac{P(S=t\mid X=x)P(S=o)}{P(S=o\mid X=x)P(S=t)}$, both estimation principles have the same limit (sufficient condition). Since $P(S=t\mid X=x) = \dfrac{f_{X,S}(x,t)}{f_X(x)} = \dfrac{f_{X|S}(x,t)P(S=t)}{f_X(x)}$ and $P(S=o\mid X=x) = \dfrac{f_{X|S}(x,o)P(S=o)}{f_X(x)}$, we obtain the second representation of the weights as $w = w(x) = \dfrac{f_{X|S}(x,t)}{f_{X|S}(x,o)}$ shown in Theorem 1a). q.e.d.

The proof of Theorem 1b) proceeds along the same lines as the previous one, but without the explicit conditioning on X that is not necessary in this case. Thus, we get:

$$F_{Y|S}^{M(w)}(y,t) = \underset{W|S=t}{E}\left[F_{Y|W,S}(y,w,o)\right] \overset{Bayes\ Law}{=} \underset{W|S=o}{E}\left[\frac{P(S=t\mid W=w)P(S=o)}{P(S=o\mid W=w)P(S=t)}F_{Y|W,S}(y,w,o)\right].$$

Therefore, if $w = \frac{P(S=t\mid W=w)P(S=o)}{P(S=o\mid W=w)P(S=t)} = \frac{f_{W|S}(w,t)}{f_{W|S}(w,o)}$, both estimation principles have the same limit. q.e.d.

A.2 Proof of Theorem 3

In the proof of this theorem the law of total probability will be frequently applied, i.e. $\frac{f_{A,B}(a,b)}{f_{A|B}(a,b)} = f_B(b) = \frac{f_{A,B}(a',b)}{f_{A'|B}(a',b)}$, with A, B being random variables. From this property, we get three conditions that are helpful in the proof of this theorem (as well as of Theorem 2b'):

$$F_{Y|W,S}(y,w,t) = \frac{P(S=t\mid Y=y,W=w)}{P(S=o\mid Y=y,W=w)}\frac{P(S=o\mid W=w)}{P(S=t\mid W=w)}F_{Y|W,S}(y,w,o); \tag{A.1}$$

$$f_{W|S}(w,t) = \frac{P(S=t\mid W=w)P(S=o)}{P(S=o\mid W=w)P(S=t)}f_{W|S}(w,o); \tag{A.2}$$

$$f_{W|S}(w,o) = \frac{P(S=o\mid W=w)P(S=t)}{P(S=t\mid W=w)P(S=o)}f_{W|S}(w,t). \tag{A.3}$$

Using those properties, the proof of the first part of Theorem 3 is direct:

$$\begin{aligned}
&\underset{W|S=t}{E}F_{Y|W,S}(y,w,t) - \underset{W|S=o}{E}wF_{Y|W,S}(y,w,o) = \\
&\overset{(A.3)}{=} \underset{W|S=o}{E}\left[\frac{P(S=t\mid W=w)P(S=o)}{P(S=o\mid W=w)P(S=t)}F_{Y|W,S}(y,w,t)\right] - \underset{W|S=o}{E}wF_{Y|W,S}(y,w,o) = \\
&\overset{(A.1)}{=} \underset{W|S=o}{E}\left[\frac{P(S=o)}{P(S=t)}\frac{P(S=t\mid Y=y,W=w)}{P(S=o\mid Y=y,W=w)}F_{Y|W,S}(y,w,o)\right] - \underset{W|S=o}{E}wF_{Y|W,S}(y,w,o) = \\
&= \underset{W|S=o}{E}\left\{\left[\frac{P(S=o)}{P(S=t)}\frac{P(S=t\mid Y=y,W=w)}{P(S=o\mid Y=y,W=w)} - w\right]F_{Y|W,S}(y,w,o)\right\} \overset{!}{=} 0.
\end{aligned}$$

We get the second part of Theorem 3 by combining the last line above with property (A.2):

$$\underset{W|S=o}{E}\left\{\left[\frac{P(S=o)}{P(S=t)}\frac{P(S=t\,|\,Y=y,W=w)}{P(S=o\,|\,Y=y,W=w)}-w\right]F_{Y|W,S}(y,w,o)\right\}=$$

$$\overset{(A.2)}{=}\underset{W|S=t}{E}\left\{\frac{P(S=o\,|\,W=w)P(S=t)}{P(S=t\,|\,W=w)P(S=o)}\left[\frac{P(S=o)}{P(S=t)}\frac{P(S=t\,|\,Y=y,W=w)}{P(S=o\,|\,Y=y,W=w)}-w\right]F_{Y|W,S}(y,w,o)\right\}=$$

$$=\underset{W|S=t}{E}\left\{\frac{P(S=o\,|\,W=w)}{P(S=t\,|\,W=w)}\left[\frac{P(S=t\,|\,Y=y,W=w)}{P(S=o\,|\,Y=y,W=w)}-w\frac{P(S=t)}{P(S=o)}\right]F_{Y|W,S}(y,w,o)\right\}\overset{!}{=}0.$$

q.e.d.