DISCUSSION PAPER SERIES

高麗大學校 經濟研究所

THE INSTITUTE OF ECONOMIC RESEARCH
KOREA UNIVERSITY

『Sensitivity Analysis of Job-Training Effects on Reemployment for Korean Women』

Myoung-jae Lee and Sang-jun Lee

Discussion Paper No. 07-20 (July 2007)

The Institute of Economic Research Korea University
Anam-dong, Sungbuk-ku,
Seoul, 136-701, Korea
Tel: (82-2) 3290-1632 Fax: (82-2) 928-4948

SENSITIVITY ANALYSIS OF JOB-TRAINING EFFECTS ON REEMPLOYMENT FOR KOREAN WOMEN

(July 9, 2007)

Myoung-jae Lee
Department of Economics, Korea University
Anam-dong, Sungbuk-ku, Seoul 136-701
Republic of Korea
myoungjae@korea.ac.kr; fax: 82-2-926-3601

Sang-jun Lee
Korea Research Institute for Vocational Education and Training
15-1, Chongdam 2-dong, Kangnam-gu, Seoul, South Korea
sjlee@krivet.re.kr; fax: 82-2-3485-5059

The main difficulty in treatment effect analysis with matching is accounting for unobserved differences (i.e., selection problem) between the treatment and control groups, because matching assumes no such differences. The traditional way to tackle the difficulty has been 'control function' approaches with selection correction terms. This paper examines relatively new approaches: sensitivity analyses—sensitivity to unobservables—in Rosenbaum (1987), Gastwirth et al. (1998), and Lee (2004). These sensitivity analyses are applied to the data used in Lee and Lee (2005) to see how the assumption of no unobserved difference in matching affects the findings in Lee and Lee, to compare how the different sensitivity analyses perform, and to relate the 'sensitivity parameters' in the different sensitivity analyses to one another. We find (i) the conclusions in Lee and Lee are weakened in the sense that only the 'strong' ones survive, (ii) the sensitivity analysis in Rosenbaum (1987) is too conservative (and inferior to Gastwirth et al.'s), and (iii) Gastwirth et al.'s (1998) and Lee's (2004) approaches agree on some findings to be insensitive, but the two approaches also disagree on some other findings. We also look for 'comparable values' for the sensitivity parameters such that the resulting sensitivity findings are comparable across the different sensitivity analyses.

Key Words: matching, sample selection, sensitivity analysis, job training.

1 Introduction

1.1 Motivation

In finding the effect of a treatment d on a response variable y of interest, matching has been the main method comparing the treatment and control (or 'comparison') groups matched on some observed covariates x (or a function of x). The assumption justifying matching is that there is no systematic difference in unobserved variables v that causes a bias to the matching estimator—the assumption is often called 'selection-on-observables'. But when the treated subjects systematically differ from the control subjects in x, it is natural to raise the question whether they might systematically differ in v as well, which is called 'selection-on-unobservables'.

Different approaches have been used in the treatment effect literature to relax the selection-on-observable assumption: instrumental variables, 'control functions', and 'bounding approaches'. But plausible instruments are hard to find in practice; control function approaches as in Heckman et al. (2003) are heavily parametric; bounding approaches as in Manski (2003), which are nonparametric, tend to yield too large bounds to be useful.

An alternative to these conventional approaches is *sensitivity analysis* where the presence of the bias-causing unobserved variables v is indexed by a 'sensitivity parameter', say γ, with $\gamma = 0$ (or $\gamma = 1$, depending on the parametrization) meaning no unobserved variables. Initially, a finding is derived under selection-on-observables $\gamma = 0$, and then the assumption $\gamma = 0$ is relaxed to assess how the initial finding is affected. If the initial finding changes "much" as γ deviates from zero only "slightly", then the finding is deemed sensitive; otherwise, the finding is insensitive. The aforementioned bounding approaches bear some resemblance to this type of sensitivity analysis, but a good way to differentiate the former from the latter seems to be that the former is global while the latter is local (to $\gamma = 0$).

Although we may have a good idea on how big is big for changes in y, the vexing question in any sensitivity analysis is how big is big for $|\gamma|$. But this arbitrariness of $|\gamma|$ differs little from that in choosing the level in tests. As more 'track records' get built up in sensitivity analyses, there may emerge some consensus over the size of $|\gamma|$. Sensitivity analysis as presented in this paper may become another weapon in researchers' arsenal in allowing for selection on unobservables. For this to happen, as just mentioned, track records are needed, and this paper takes a step toward that direction. Although applied studies routinely use various forms of sensitivity analysis to check out the model assumptions, sensitivity analysis for identifying assumptions—in our case, sensitivity to unobservables—is not often seen.

For treatment effect analysis with matching, various sensitivity analyses have appeared in the statistics literature (see Rosenbaum (2002)) depending on (i) which matching method is used (ii) for what type of response variable (iii) with which test statistic. Applied in this paper are two sensitivity analyses in Rosenbaum (1987) and Gastwirth et al. (1998) to (i)' pair matching method (ii)' for binary response (iii)' with 'McNemar statistic' which counts how many successes are associated with the treated units in the pairs. Although the setup (i)' to (iii)' may look restrictive, they constitute a base case, because pair matching can be extended to multiple matching, a binary response can be generalized to non-binary responses, and McNemar statistic is a natural test statistic for pair-matched samples with binary responses.

Outside the statistical literature, not many sensitivity analysis studies have appeared so far. Aakvik (2001) assesses the effects of a vocational rehabilitation program with a binary response; Aakvik groups the data into 12 strata based on the 'propensity score' $P(d = 1|x)$ and uses 'Mantel-Haenszel' statistic and a sensitivity analysis in Rosenbaum (2002, p.130), which have been also used by Hujer et al. (2004), Caliendo et al. (2005), and Hujer and Thomsen (2006) for effects of German job-creation schemes. While these papers only apply existing methods, Imbens (2003) presents a new parametric sensitivity analysis for normally distributed continuous responses and a binary treatment derived from the logistic distribution. Imbens' study builds on the early work of Rosenbaum and Rubin (1983b) that also motivated another sensitivity analysis in Altonji et al. (2005), who use the degree of selection on observables as a guide on the degree of selection on unobservables. There is no doubt that the number of papers on sensitivity analysis will be increasing in future.

A problem with Rosenbaum and Rubin (1983b), Imbens (2003) and Altonji et al. (2005) is that their sensitivity analyses require parametric assumptions that are not necessary for estimating the treatment effect per se. Ichino et al. (2007) suggest the following sensitivity analysis for binary treatment and response without imposing extra parametric assumptions such as normal or logistic distributions. Suppose a significant treatment effect has been found using a matching scheme. To see if this is due to an unobserved binary covariate, simulate a binary covariate v for each individual subject to some specified four cell proportions $p \equiv (p_{00}, p_{01}, p_{10}, p_{11})'$—there are $2 \times 2 = 4$ cells from the binary treatment and response—

and use the artificial v along with the actual covariates in matching to obtain a "pseudo" treatment effect. This simulation is repeated many times to yield many pseudo treatment effects, from which their average and variance can be estimated. If the standardized pseudo average treatment effect is insignificant, then the initial significant treatment effect might have been due to not using v. The sensitivity parameter here is p, and if it takes an extreme (i.e., implausible) value of p to negate the initial significant effect, then the initial finding is deemed insensitive. The four cell probabilities for the observed binary covariates may be taken as plausible values of p. With the subscripts in p_{jk} corresponding to $d = j$ and $y = k$, Ichino et al. further narrowed down the sensitivity parameters into two: $p_{01} - p_{00}$ (the v effect on the untreated response) and $p_{1.} - p_{0.}$ (the v effect on d) where $p_{j.} \equiv P(v = 1|d = j)$.

Different sensitivity analyses require different assumptions and thus detect departures from the no-selection-bias assumption in different directions. Insensitivity of an empirical finding to multiple sensitivity analyses makes the finding more credible. This motivates applying multiple sensitivity analyses jointly to the same data. This situation is analogous to using different instruments for the same endogenous regressor. While instrumental variable estimator (IVE) with a single instrument may not be so credible because the instrument may be "faulty", if IVE's with different instruments all point to the same direction, the finding there becomes more credible, as noted in Imbens and Rosenbaum (2005). Applying and comparing different sensitivity analyses together, we can also relate the sensitivity parameters in different sensitivity analyses to one another, which is helpful in choosing the size of the sensitivity parameters.

In analyzing the effects on unemployment duration of classroom-teaching-type job-trainings in Korea, Lee and Lee (2005, LL from now on) conclude that, unfortunately, the trainings have negative effects, lengthening the duration that includes the training duration. More specifically, the trainings were not cost-effective in that they took too much time (about four months) "locking in" the trainees during the training span, compared with the time they took to place the trainees afterwards (about three months shortened compared with the non-trainees). Despite this negative finding, however, some job trainings indeed had positive effects: finance/insurance and information/communication. LL's study is based on selection-on-observables. As mentioned above, we will apply the two sensitivity analyses in Rosenbaum (1987) and Gastwirth et al. (1998) to the same data in LL and see to what extent the findings in LL survive the introduction of selection on *un*observables. We will also apply a sensitivity analysis in Lee (2004), which differs much from the aforementioned sensitivity analyses and is more broadly applicable regardless of the response variable type.

1.2 Basic Framework

Before proceeding further, we present the basic framework of treatment effect analysis where a job-training is the treatment and the indicator function for finding a job or not by the study-ending date is the response—this is the case for our empirical analysis later. Let y_{ji} be the 'potential response' when the treatment $d_i = j$, $j = 0, 1$, is "exogenously given to" individual i, $i = 1, ..., N$; the observed response is $y_i = d_i y_{1i} + (1 - d_i) y_{0i}$. Rubin (1974) seems to be the first to formalize this potential versus observed response setup, although the idea dates further back into the past. Assuming *iid* across $i = 1, ..., N$, we will often omit the subscript i from now on.

The usual treatment effect of interest is the 'mean effect' $E(y_1 - y_0)$. If $y_j, j = 0, 1$, are mean-independent of d (i.e., if $E(y_j|d) = E(y_j)$), then the mean effect is the same as the group mean difference $E(y|d = 1) - E(y|d = 0)$; for a binary response as in our data, mean independence is the same as independence. For the 'mean effect on the treated' $E(y_1 - y_0|d = 1)$ to be identified by the group mean difference, the mean independence of only y_0 from d is sufficient as shown below.

The mean independence holds if d is randomized as in clinical trials. But, in observational data as ours, d is not randomized but self-selected by the individuals. Thus, the treatment group ('T group') can be systematically different at a pre-treatment stage from the control group ('C group') in terms of x or v. To make matters worse, 37% of the trainees in our data dropped out ('D group'). The D group was not used at all in LL and will be used only occasionally as part of the C group in this paper, because it is not clear how to handle the D group. Particularly those who dropped out because they had found a job are troublesome (dropout reasons are not known in the data), because the dropout decision was affected by the response variable. These considerations make the mean independence assumption questionable, which is why sensitivity analysis is called for.

The differences in x and v between the T and C groups should be eliminated. If not, the difference in x may cause an 'overt bias', and the difference in v may cause a 'hidden bias'; these terminologies are taken from Rosenbaum (2002). Most effort in treatment effect analysis with observational data is spent on eliminating these biases. See Heckman et al.

(1999), Angrist and Krueger (1999), Rosenbaum (2002), and Lee (2005) among others for more on treatment effect in general.

The overt bias can be removed by conditioning on x: instead of the unconditional group mean difference, examine the conditional group mean difference

$$E(y|x, d = 1) - E(y|x, d = 0) = E(y_1|x, d = 1) - E(y_0|x, d = 0) = E(y_1|x) - E(y_0|x)$$

where the last equality holds if

$$y_0 \text{ and } y_1 \text{ are (mean-) independent of } d \text{ given } x.$$

This is a selection-on-observables assumption that assumes away hidden bias. Once $E(y_1 - y_0|x)$ is found, x can be integrated out to yield a 'marginal effect'.

For 'the mean treatment effect on the treated' $E(y_1 - y_0|d = 1)$, we just need the mean independence of y_0 from d given x, which implies

$$E(y|x, d = 1) - E(y|x, d = 0) = E(y_1|x, d = 1) - E(y_0|x, d = 0)$$
$$= E(y_1|x, d = 1) - E(y_0|x, d = 1) = E(y_1 - y_0|x, d = 1).$$

Integrating out x using the distribution $F(x|d = 1)$ of $x|(d = 1)$ yields the marginal effect on the treated. Heckman et al. (1998) decompose the bias $E(y_0|d = 1) - E(y_0|d = 0)$ into three terms: the first is due to the non-overlapping supports between $F(x|d = 1)$ and $F(x|d = 0)$, the second due to $F(x|d = 1) \neq F(x|d = 0)$ on the overlapping support, and the third due to $E(y_0|x, d = 1) \neq E(y_0|x, d = 0)$. Among these three, we are going after the third: the bias caused by not being able to control for the unobservables.

Suppose selection-on-unobservables holds in the sense that

$$y_0 \text{ and } y_1 \text{ are (mean-) independent of } d \text{ given } x \text{ and } v.$$

Here, d can be determined by three factors: x, v, and some random components, say ε, unrelated to y_0 and y_1 given x and v. A sensitivity analysis can be conducted by allowing v to affect y_0, y_1, and d. For instance, consider

$$d_i = 1[x_i'\alpha + \gamma v_i + \varepsilon_i > 0] \quad \text{and} \quad y_{ji} = 1[x_i'\beta_j + \delta v_i + u_{ji} > 0] \tag{1.1}$$

where $1[A] = 1$ if A holds and 0 otherwise, α and β_j are parameters, and the treatment equation error term ε_i is independent of the potential response error u_{ji} given x_i and v_i,

$j = 0, 1$. The presence of v is indexed by the sensitivity parameters γ and δ. Although v cannot be controlled for, its presence can be accounted for with a sensitivity analysis.

Selection-on-observable assumption cannot be tested with observational data unless there are valid instruments essentially amounting to randomization indicators. The plausibility of the assumption depends ultimately on how detailed the observational data is in terms of the characteristics of the individuals, the treatments under consideration, and the process of determining how the treatments get assigned. The list of the available covariates is not extensive in LL, which thus casts doubts on the selection-on-observable assumption. In the LL data, there are many potentially relevant unobservables one could think of: motivation of the individuals, marital status (and her husband work/income status if married), the family characteristics, case worker discretion in choosing the right training type, and so on. Hence sensitivity analysis is called for, and it is a logical step forward from LL.

The rest of this paper is organized as follows. Section 2 reviews briefly the three sensitivity analyses. Section 3 introduces the data. Section 4 presents the empirical findings. Finally, Section 5 concludes.

2 Review of Three Sensitivity Analyses

2.1 Unobserved Confounder Affecting Treatment

Suppose that v is the only unobserved confounder. For v to cause a hidden bias, v should affect both d and y_j equation as (1.1) illustrates. Because v affecting d is necessary for hidden bias, regardless of whether v affects y_j or not, a sensitivity analysis is possible using only γ parametrizing the presence of v in the d equation. Rosenbaum (1987) presents such a sensitivity analysis.

Rosenbaum (2002) shows that, for two matched subjects with (x, v_1) and (x, v_2),

$$\begin{aligned} & 0 \le v \le 1 \ \text{ and } \ \varepsilon \text{ as in (1.1) follows logistic distribution independently of } x \text{ and } v \\ \text{iff} \quad & \Gamma^{-1} \le \frac{P(d=1|x, v_1)/P(d=0|x, v_1)}{P(d=1|x, v_2)/P(d=0|x, v_2)} \le \Gamma, \quad \text{where } \Gamma \equiv \exp(\gamma) \ge 1 \end{aligned} \tag{2.1}$$

and conducts a sensitivity analysis with $\Gamma \ge 1 \iff \gamma \ge 0$. $\Gamma = 1$ corresponds to no selection bias, in which case the odds ratio of receiving the treatment becomes one: the two units are equally likely to get treated.

Consider only the discordant pairs (i.e., the responses differ within each matched pair), say, n many with $x_{s1}, x_{s2},\ v_{s1}, v_{s2},\ d_{s1}, d_{s2},\ y_{s1}, y_{s2},\ s = 1, ..., n$; order each pair such that $y_{s1} = 0$ and $y_{s2} = 1$. By construction, $x_{s1} = x_{s2} \equiv x_s$, $d_{s1} \neq d_{s2}$, and $y_{s1} = 0$ and $y_{s2} = 1$. *McNemar's statistic* $M \equiv \sum_{s=1}^{n} d_{s2}$ is the number of treated subjects with $y = 1$ among n-many discordant pairs. Let m be the realized value of M, and assume $m/n > 0.5$ without loss of generality, for the definition of y can be reversed if $m/n \leq 0.5$. The key entity is

$$P(d_{s2} = 1 | y_{s1} = 0, y_{s2} = 1, d_{s1} \neq d_{s2},\ x_s, v_{s1}, v_{s2}).$$

Suppose that the treatment has no effect to yield

$$P(d_{s2} = 1 | y_{s1} = 0, y_{s2} = 1, d_{s1} \neq d_{s2},\ x_s, v_{s1}, v_{s2}) = P(d_{s2} = 1 | d_{s1} \neq d_{s2},\ x_s, v_{s1}, v_{s2}).$$

If $\gamma = 0$ or $v_{s1} = v_{s2}$, then $P(d_{s2} = 1 | d_{s1} \neq d_{s2},\ x_s, v_{s1}, v_{s2})$ becomes 0.5, which results in $M \sim B(n, 0.5)$. If $\gamma > 0$ and $v_{s1} \neq v_{s2}$ for some s, then M no longer follows $B(n, 0.5)$ under no effect, because d_{s2}'s are independent but follow different binary distributions across $s = 1, ..., n$. In this case, Rosenbaum (1987) shows

$$P(M \geq m) \leq \sum_{a=m}^{n} \binom{n}{a} (p^+)^a (1 - p^+)^{n-a}, \quad \text{where } p^+ \equiv \frac{\Gamma}{1 + \Gamma} \geq 0.5.$$

With this, we can get an upper bound on the p-value for the one-sided test rejecting the H_0 of no effect when m is large.

With a large n, the upper bound of the p-value $P(M \geq m)$ is

$$P(B(n, p^+) \geq m) \simeq P(N(0, 1) \geq \frac{m - np^+}{\sqrt{np^+(1 - p^+)}}) \ \ (\geq 0.5 \ \text{ if } p^+ \geq m/n);$$

the upper bound is informative only when $p^+ < m/n$. Compute $\sum_{a=m}^{n} \binom{n}{a} (p^+)^a (1 - p^+)^{n-a}$ over $0.5 \leq p^+ < m/n$, and see at which values of p^+ (i.e., Γ) the upper bound crosses the nominal level 5%.

The main question is how to interpret Γ. Suppose the upper bound of the p-value is 0.024 with $\Gamma = 2$ (for $\Gamma < 2$, the upper bound is even smaller) and 0.10 with $\Gamma = 3$. Here, the decision of rejecting H_0 at level 5% gets reversed as Γ changes from 2 to 3. The key question is whether $\Gamma = 3$ is a plausible value or not: the unobserved confounder v should be able to produce a three-fold increase in the odds ratio of getting treated between the two subjects in the same pair as (2.1) shows. That is, if v were observed, then one would be able to identify the subject whose odds of receiving the treatment is three times greater than the other subject's odds. If most relevant variables for treatment decision have been already included in x and if it is not easy to think of a variable resulting in that much difference in the odds ratio, then $\Gamma = 3$ is an implausibly large value. The fact that it takes such a big number to reverse the initial conclusion makes the initial conclusion of rejecting H_0 insensitive to the hidden bias. In our job-training data, in view of $0 \leq v \leq 1$ in (2.1), the question would be whether there is any unobserved variable v that makes one subject's odds with $v = 1$ three times greater the other subject's odd with $v = 0$ when the two subjects' x is the same.

2.2 Unobserved Confounder Affecting Treatment and Response

The preceding sensitivity analysis can be too conservative, because there should be no hidden bias if v does not affect (y_0, y_1) even when v affects d; ignoring this leads to a bound on the p-value even when the bound is in fact unwarranted. Gastwirth et al. (1998) generalize the sensitivity analysis of Rosenbaum (1987) and allow v to affect both (y_0, y_1) and d at the price of one additional sensitivity parameter and the logistic distribution assumption below. Rosenbaum and Rubin (1983b) considered v affecting both treatment and response for the first time, but their approach contains too many sensitivity parameters.

Instead of (2.1), assume now, for $i = 1, 2$,

$$P(d_{si} = 1|x, v_{si}) = \frac{\exp\{\alpha_s + f_d(x) + \gamma v_{si}\}}{1 + \exp\{\alpha_s + f_d(x) + \gamma v_{si}\}}, \quad P(y_{si} = 1|x, v_{si}) = \frac{\exp\{\kappa_s + f_y(x) + \delta v_{si}\}}{1 + \exp\{\kappa_s + f_y(x) + \delta v_{si}\}} \tag{2.2}$$

for some constants α_s and κ_s, unknown functions f_d and f_y, and $0 \leq v_{si} \leq 1 \ \forall s, i$. If $\gamma\delta \neq 0$, then v_{si} affects both the treatment and response, causing a hidden bias; otherwise, there is no hidden bias. Differently from (2.1), the assumption of the logistic forms and $0 \leq v_{si} \leq 1$ do not follow from the bounds on the odds ratio. The bound $0 \leq v_{si} \leq 1$ is not so restrictive, however, because v_{si} can be re-scaled with γ and δ and re-centered with α_s and κ_s.

Define

$$\begin{aligned}
\Gamma &\equiv \exp(|\gamma|), \quad \Delta \equiv \exp(|\delta|), \\
\zeta(\Gamma, \Delta) &\equiv \frac{\Gamma}{1+\Gamma}\frac{\Delta}{1+\Delta} + (1 - \frac{\Gamma}{1+\Gamma})(1 - \frac{\Delta}{1+\Delta}) \\
&= 2(\frac{\Gamma}{1+\Gamma} - 0.5)(\frac{\Delta}{1+\Delta} - 0.5) + 0.5 \geq 0.5 \quad \text{as} \quad \frac{\Gamma}{1+\Gamma}, \frac{\Delta}{1+\Delta} \geq 0.5.
\end{aligned}$$

When $\gamma\delta = 0$, $M \sim B(n, 0.5)$ under no effect; when $\gamma\delta \neq 0$, Proposition 1 in Gastwirth et al.

(1998) shows that, under no effect,

$$P(M \geq m) \leq \sum_{a=m}^{n} \binom{n}{a} \zeta(\Gamma, \Delta)^a \{1 - \zeta(\Gamma, \Delta)\}^{n-a}.$$

Compute the right-hand side over $0.5 \leq \zeta < m/n$. This is almost the same as $\sum_{a=m}^{n} \binom{n}{a} (p^+)^a (1-p^+)^{n-a}$ in the previous subsection. The only difference is that Δ as well as Γ appear in ζ whereas only Γ appears in p^+.

2.3 Average Ratios of Biased to True Effects

Lee (2004) adopts a geometric-average-based marginal effect

$$\left\{ \prod_i \frac{E(y_1|x_i)}{E(y_0|x_i)} \right\}^{1/N} - 1 = \exp\{\frac{1}{N} \sum_i \ln \frac{E(y_1|x_i)}{E(y_0|x_i)}\} - 1$$

motivated by the *conditional proportional effect*

$$\frac{E(y_1 - y_0|x)}{E(y_0|x)} = \frac{E(y_1|x)}{E(y_0|x)} - 1$$

which shows the proportional effect relative to the baseline $E(y_0|x)$.

Rewrite the conditional proportional effect plus one as

$$\frac{E(y_1|x)}{E(y_0|x)} = \frac{E(y|x, d=1)}{E(y|x, d=0)} A(x) \quad \text{where } A(x) \equiv \frac{E(y|x, d=0)}{E(y_0|x)} \cdot \left\{ \frac{E(y|x, d=1)}{E(y_1|x)} \right\}^{-1}.$$

The geometric-average-based marginal effect is then

$$\exp\{\frac{1}{N} \sum_i \ln R(x_i) + \frac{1}{N} \sum_i \ln A(x_i)\} - 1 = B \exp\{\frac{1}{N} \sum_i \ln R(x_i)\} - 1$$

$$\text{where } \; R(x_i) \equiv \frac{E(y|x_i, d=1)}{E(y|x_i, d=0)} \quad \text{and} \quad B \equiv \exp\{\frac{1}{N} \sum_i \ln A(x_i)\}.$$

This marginal effect is identified if $B = 1$; B is the geometric average of $A(x_i)$'s, $i = 1, ..., N$. When $B \neq 1$ due to selection on unobservables, for a chosen range of B, we can get the corresponding range for the marginal effect, which leads to a sensitivity analysis with B as the sensitivity parameter.

There are reasons to believe that B would not deviate too much from one. To see this, examining $A(x)$ closely, $A(x)$ consists of two ratios:

1. If both ratios in $A(x_i)$ are one (i.e., if selection-on-observables holds), then $A(x_i)$ is one.

2. Even if the ratios differ from one in $A(x_i)$ (i.e., even if there are hidden biases), so long as they are equal (i.e., so long as the bias magnitudes are the same in the two ratios), then $A(x_i)$ is still one.

3. Even if $A(x_i) \neq 1$ for some i, B can still be close to one, for B is an average of $A(x_i)$'s.

Under these conditions, we need to search only in a small neighborhood of $B = 1$.

Specifically, a *nonparametric sensitivity analysis* can be done as follows. *First*, estimate $R(x_i)$ with a kernel estimator (other nonparametric estimators can be used as well)

$$R_N(x_i) \equiv \frac{\sum_{j,j\neq i} K(\frac{x_j - x_i}{h}) d_j y_j}{\sum_{j,j\neq i} K(\frac{x_j - x_i}{h}) d_j} \cdot \left\{ \frac{\sum_{j,j\neq i} K(\frac{x_j - x_i}{h})(1-d_j) y_j}{\sum_{j,j\neq i} K(\frac{x_j - x_i}{h})(1-d_j)} \right\}^{-1}$$

where K is a kernel and h is a bandwidth converging to zero as $N \to \infty$. *Second*, obtain the marginal effect $\exp\{N^{-1}\sum_i \ln R_N(x_i)\} \;-1$. *Third*, change B around 1 to see how

$$B \cdot \exp\{\frac{1}{N}\sum_i \ln R_N(x_i)\} \;-1 \tag{2.3}$$

changes. If this changes much (e.g. going from + to − as B changes only slightly), then the proportional treatment effect is sensitive to the assumption of selection on observables. If the dimension of x is high, then following the idea in Rosenbaum and Rubin (1983a), one may use the propensity score $P(d = 1|x)$ instead of x in the nonparametric estimation to reduce the smoothing dimension to one. In practice, $P(d = 1|x)$ can be estimated by probit or logit.

The sensitivity analyses in the preceding two subsections are 'of structural form' in the sense that they are explicit on how the single unobserved confounder v may cause a hidden bias (by influencing the treatment or the response), whereas (2.3) is 'of reduced form' in the sense that such a route for hidden bias is not explicit. Consequently, a disadvantage of (2.3) is that it is less informative on how the unobserved confounder operates. But the advantage of (2.3) is that it needs hardly any assumption; e.g., (2.3) imposes no restriction on the number of unobserved confounders, whereas the two preceding sensitivity analyses allow only one. Essentially what we have is a trade-off of 'structure' and 'generality'. As more structure is imposed (one unobserved confounder v), more detailed information can be gathered (how v affects the treatment and response) at the cost of generality.

To be more specific on the sensitivity of (2.3), suppose our concern is the 'qualitative conclusion'—the effect being positive or negative. In this case, whether (2.3) crosses zero or not as B changes around 1 is the main concern. If the term $\exp\{N^{-1}\sum_i \ln R_N(x_i)\}$ next to B

is sufficiently away from one—say 2 or 0—then (2.3) does not cross zero easily as B deviates from 1. To allow for sampling error, suppose $\exp\{N^{-1}\sum_i \ln R_N(x_i)\}$ has a confidence interval (G_L, G_U), which gives the confidence interval (G_U^{-1}, G_L^{-1}) for $[\exp\{N^{-1}\sum_i \ln R_N(x_i)\}]^{-1}$. Then, because the value of B yielding (2.3) $= 0$ is $[\exp\{N^{-1}\sum_i \ln R_N(x_i)\}]^{-1}$, the initial finding under $B = 1$ is *qualitatively insensitive* if

$$(G_U^{-1}, G_L^{-1}) \notin (1-\xi, 1+\xi) \quad \text{for a small constant } \xi > 0. \tag{2.4}$$

Instead of B, ξ can be taken as the sensitivity parameter. The three reasons cited above would confine ξ to a small positive number, say 0.1, as the next section illustrates. Lee (2004) used a bootstrap to obtain (G_L, G_U).

3 Data

The data consist of two files from the Center for Employment Information in the Department of Labor of South Korea: a job-training file for the T and D groups and an unemployment-insurance file for the C group. Here, the C group are the unemployed who chose to receive unemployment insurance benefit (UIB) instead of a job training. That is, the C group is not drawn from the entire unemployed population. This is because the job-training file includes only women with some records on unemployment insurance premium payment.

Among the women who became unemployed in 1999, those who took a job-training to complete it or drop out of it by the end of 1999 or those who received UIB which ended by the end of 1999 are in the data set. This means that inclusion in the data set depends on the outcome (i.e., other than for the dropouts, this means not being reemployed until the end of the training/UIB). This problem is, however, unlikely to distort our treatment effect analysis much, because training and UIB durations are similar across the T and C groups: the mean (median) training duration is 123 (120) days and the mean (median) UIB duration is 126 (133) days. The included individuals are then followed up until the end of March 2000; this rather short follow-up resulted in a high censoring percentage (71%).

Although an unemployed woman can choose the treatment to an extent, she has to meet a criterion to be eligible for UIB: paying for the insurance at least for 0.5 year is required. For those ineligible women, the selection problem (job-training or not) is less worrisome, as it is highly likely that most unemployed women prefer "free and easy" UIB to a job-training.

Voluntary quitters are not eligible for UIB although there are exceptions. While the first selection of job-training or not is not so worrisome, a second selection problem can occur in choosing the type of job training which is determined by the case worker discussing the matter with the unemployed.

The duration of a job-training varies over 2.5 to 5 months across training types. The trainees receive some compensation, the amount of which varies. Roughly, the training compensation is about \$400 per month, and the UIB is the minimum of about \$1,000 and 50% of the last workplace monthly wage. Eight types of job-trainings are included in the data: textile, machine/equipment, information/communication, industrial application, service industry, clerical/administration, finance/insurance, and health-related. The definition of service industry is not clear in the data source, but it includes low-paying jobs such as restaurant servers and cooking assistants.

A person is recorded as a dropout if her final training compensation date precedes her full training duration. The training compensation is paid at the end of each month, and to receive the training compensation, it is required to attend the training sessions at least two thirds of the month. Once a trainee fails to receive a compensation, she is no longer eligible for the training. In the data, there is no information on dropout reasons, but talks with case workers suggest that, more often than not, dropout occurs due to a new job found. Dropout possibility adds a third selection problem. The multiple selections suggest that Lee's (2004) sensitivity analysis may be better suited than other sensitivity analyses that assume only one unobserved confounder.

The initial D and T group size was 8091 = 3024(D) + 5067(T), and the C group size was 47120. Among the initial D and T groups, 84 women from the D group and 36 from the T group were removed, either because they were in agriculture/fishery or because they had missing entries in the covariates. Among the initial C group, 60 were removed for the same reasons. Finally in our working sample, N_c, N_d, and N_t (the sample size for the C, D, and T groups) are, respectively, 47060, 2940, and 5031.

To avoid the high censoring problem and also because sensitivity analysis is yet to be developed for censored responses, we use $y = 1[\textit{employed by 31/3/2000}]$ as the response variable instead of the observed unemployment duration. In LL, matching was done also for this binary response. The information loss in using the binary response instead of the censored duration would not be too much, for the censoring percentage is high. If we compare two women—one treated and one control—matched on the quit/fired and study-entering dates as well as on the other covariates, the two women searched for jobs equally long before joining the study, and since they entered the data at the (almost) same time, we can compare them with $y = 1[\textit{employed by 31/3/2000}]$.

TABLE 1: Descriptive Statistics for Three Groups

	T		D		C	
Variables	mean (med)	SD	mean (med)	SD	mean (med)	SD
age (years)	27.8 (27)	5.57	26.8 (26)	5.06	35.0 (30)	10.8
schooling (years)	13.1 (12)	1.76	13.0 (12)	1.73	12.1 (12)	2.50
job: executive	0.003	0.05	0.01	0.07	0.01	0.10
job: professional	0.03	0.18	0.04	0.21	0.04	0.20
job: semi-professional	0.06	0.24	0.06	0.23	0.06	0.23
job: clerical	0.46	0.50	0.45	0.50	0.51	0.50
job: service/sales	0.14	0.35	0.14	0.35	0.10	0.30
job: mechanic	0.16	0.36	0.16	0.36	0.10	0.31
job: menial labor	0.12	0.33	0.12	0.33	0.17	0.37
reason: self-employed	0.48	0.50	0.50	0.50	0.02	0.15
reason: marriage/baby	0.16	0.37	0.14	0.35	0.01	0.12
reason: injured/old	0.04	0.18	0.04	0.19	0.03	0.17
reason: personal	0.16	0.36	0.14	0.35	0.004	0.06
ex-firm size/1000	1.07 (0.06)	3.76	1.14 (0.06)	4.54	0.84 (0.04)	3.28
tenure at ex-firm (years)	1.79 (0.96)	2.21	1.57 (0.74)	2.06	2.47 (2.85)	1.32
study-entering date	166 (165)	66.9	149 (152)	61.8	154 (153)	71.7
quit/fired date	101 (91)	64.7	88.1 (80)	57.6	121 (112)	71.3
% employed	27.9		39.1		28.9	
# observations	5,031		2,940		47,060	

The following covariates are used: age, schooling years (edu), last-workplace job-type (machine operator plus seven categories in Table 1) and industry type (ten categories, omitted from Table 1), reason for quitting (four categories in Table 1), last workplace size in 1000 workers, and tenure for the last workplace in years. The ten industry categories omitted from Table 1 are manufacturing, retail/wholesale, real-estate rental, finance/insurance, health-related, construction, hotel/inn, transportation/warehouse, private education, and public service. Also used as covariates are the quit/fired date in the last workplace and the date of entering the study by starting a training or receiving UIB; the calendar dates are recorded with 01/01/1999 as 1 (i.e., 31/12/1999 is 365). Although not shown in Table 1, most average covariate differences across the T and C groups are statistically significant with absolute t-values often running well over 10.

In Table 1, the average age of group C is higher (35) than that of group T (28) and group D (27); this is understandable from the standard human capital theory that the trainees have more time to reap the benefit of their investment in human capital. The previous job type is similar across the three groups. In group T, 48% quit their jobs because of self-employment and 16% getting married or having a baby. The reason for quitting is not known for 16% ('reason: personal'). The row 'ex-firm size /1000' shows the size (in 1000 persons) of the last workplace; the majority are under 0.5, but due to some outliers of size more than 10, the mean and SD of ex-firm size are large. The average tenure at the last workplace is 1.79 years for group T, 1.57 for group D, and 2.47 for group C. The duration between the time points of losing the job and enrolling in a job-training program is about two months, whereas the duration between the time points of losing a job and receiving UIB is about one month.

Overall, although groups T and C differ much, there is hardly any difference between groups T and D. When probit is applied to T and C (the probit result is omitted), a pseudo R^2 is 60%: much of the training or no-training decision is explained by the observed variables. This is mainly because women eligible for UIB tend to choose no training. In contrast, the probit for dropout decision is hardly explained by any observed variable with the pseudo R^2 only 1.5% using the same probit model.

4 Empirical Findings

4.1 First Sensitivity Analysis

Table 2 shows part of the table 3 in LL for many groups: the column 'est.' and 't-value' are pair-matching treatment-effect estimates and t-values under no hidden bias. For instance, for the middle-school group, the probability of reemployment for the treated by the time the study ended is 0.175 less than that for the untreated. For each treated subject, the pair matching picks the closest control in terms of a standardized quadratic distance in x

(subject to a modification for ties as explained shortly). Once the matching is done, a pair is treated as a single observation, and the average of the response differences across the pairs and its usual t-value are shown in the column 'est' and 't-value'. The matching for Table 2 is "greedy" in the sense that each control subject is used only once at most; non-greedy matching was also applied in LL with little difference in covariate balance owing to the large control reservoir. Note that, depending on education, previous job and training status, one person may appear in several rows in Table 2. For each sub-group, all covariates except the group defining variable are used for matching.

TABLE 2: No Hidden Bias Results and Discordant Pairs n

	est.	t-value	# matched treated	n	m	$n-m$	effect
edu.: middle school	-0.175	-3.62	57	15	4	11	−
edu.: high school	-0.061	-5.40	2,057	791	333	458	−
edu.: junior college	-0.074	-3.53	715	301	127	174	−
edu.: college	-0.002	-0.10	651	292	147	145	+
job: professional	-0.047	-0.99	130	51	23	28	−
job: semi-professional	-0.051	-1.31	193	80	37	43	−
job: clerical	-0.061	-4.49	1,870	789	344	445	−
job: service/sales	-0.053	-2.44	494	168	72	96	−
job: mechanic	-0.031	-1.34	386	151	72	79	−
job: machine operator	-0.145	-2.26	43	13	5	8	−
job: menial labor	-0.069	-2.78	369	148	58	90	−
training: textile	-0.079	-1.56	90	33	11	22	−
training: machine/equip.	0.051	0.87	92	38	23	15	+
training: info./comm.	0.049	2.53	846	365	210	155	+
training: industrial appl.	-0.107	-2.82	225	90	33	57	−
training: service	-0.141	-10.2	1,256	440	130	310	−
training: clerical	-0.044	-1.93	608	251	117	134	−
training: finance/ins.	0.148	2.03	60	28	21	7	+
training: health	-0.038	-1.30	359	139	65	74	−

One concern in the matching may be that the quitting reason dummies are almost zero for group C as shown in Table 1. For instance, quitting reason 'self-employed' is only 2% for group C whereas it is 48% in group T. But group C is almost 10 times larger in size than group T, which means that there are only about 2.5 times as many subjects with quitting reason 'self-employed' in group T as in group C. By using a small 'caliper', which is a matching discrepancy threshold, not well matched treated units are discarded and the imbalance between the two groups in quitting reasons becomes much smaller. The downside of using the small caliper is, however, that quite a few treated subjects are discarded due to no closely matched control. This can be seen by comparing N_t times the proportions of the job categories in Table 1 to '# matched treated' in Table 2: about 50-80% of the treated in each group were discarded.

To gauge how well the supports of the covariates overlap across the T and C groups, although our matching did not use propensity score itself, the appendix shows propensity score frequencies for ten groups: middle school, high school, junior college, job 'clerical', job 'mechanic', job 'menial labor', training 'info./comm.', training 'industrial appl.', training 'service', and training 'finance/ins.'; as it will turn out, these 10 groups are of the main interest. The histograms show that the supports overlap fairly well except at the upper end in trainings 'info./comm. and 'service'.

The first two columns of Table 2 show that the treatment effect is negative for most groups but significantly positive for two training types: info./comm. and finance/ins. The t-value for training 'service' is particularly high, being above 10 in absolute value. The remaining columns of Table 2 show the number of matched treated subjects, the number n of discordant pairs, McNemar's statistic value m, $n - m$, and the effect sign based on whether $m/n > 0.5$ or not. The last column agrees with the sign of the first column 'est.' except for group 'college' in which $m/n \simeq 0.5$. The reason for this small discrepancy is that, when there were ties in the LL's pair matching, the average values of x and y among the tied subjects were used (thus, y could be a fraction), whereas a control is selected randomly out of the tied controls in this paper's pair matching (for y should not be a fraction for m).

In general, the greater m/n is and the greater n is, the less sensitive is the effect. For instance, training 'service' has $n = 440$ and $m = 130$, which is unlikely if no effect holds, because this means getting 130 tails in tossing a fair coin 440 times; with effect this strong, it will take a very large hidden bias to overturn the effect. Table 3 shows the results of the first sensitivity analysis with $\Gamma = \exp(\gamma)$ in (2.1). The column with $\Gamma = 1$ shows the p-values for negative/positive effects under no hidden bias. The p-values were computed with $B(n, 0.5)$ using the Sterling's formula for factorials. The entry 'NI' means "non-informative" (recall the case $p^+ \geq m/n$).

TABLE 3: Sensitivity for Treatment Assignment

	Γ=1	Γ=1.5	Γ=2	Γ=2.5
edu.: middle school	0.061	0.185	0.284	0.328
edu.: high school	0.000	NI	NI	NI
edu.: junior college	0.004	NI	NI	NI
edu.: college	0.477	NI	NI	NI
job: professional	0.289	NI	NI	NI
job: semi-professional	0.289	NI	NI	NI
job: clerical	0.000	NI	NI	NI
job: service/sales	0.038	NI	NI	NI
job: mechanic	0.313	NI	NI	NI
job: machine operator	0.297	NI	NI	NI
job: menial labor	0.005	0.170	NI	NI
training: textile	0.040	NI	NI	NI
training: machine/equip.	0.129	NI	NI	NI
training: info./comm.	0.002	NI	NI	NI
training: industrial appl.	0.007	0.154	NI	NI
training: service	0.000	0.000	0.017	NI
training: clerical	0.156	NI	NI	NI
training: finance/ins.	0.006	0.061	NI	NI
training: health	0.249	NI	NI	NI

The statistical significance based on McNemar statistic p-value at 5% level agrees mostly with that in Table 2 column 't-value'. Other than training 'service' and perhaps training 'finance/insurance', all significant results become insignificant with $\Gamma = 1.5$. When $\Gamma = 2$, only the effect for training 'service' is still significant. When $\Gamma = 2.5$, no effect is significant. In using a similar sensitivity analysis, Aakvik (2001) seems to regard a finding insensitive when it survives $\Gamma = 2$. Hujer et al. (2004), Caliendo et al. (2005), and Hujer and Thomsen

(2006) seem to adopt a lower threshold for Γ, say $\Gamma = 1.5$. As just noted, in either case, only the negative *effect for training 'service' is insensitive* in the first sensitivity analysis, with training 'finance/insurance' being nearly insensitive for $\Gamma = 1.5$ but still falling a little short.

To see whether the above results are sensitive to not using the D group, we included the D group in the C group, and re-estimated Tables 2 and 3. The results are in the appendix Table 2A and 3A, respectively. The reason for including the D group in the C group, not in the T group, is twofold. One is the median training duration of the D group is 46 days, which is 38% of the T group median training duration (120 days). The other is that Korea is a very much 'diploma/certificate-driven', rather than 'merit-driven', society: not completing a course is typically taken as never having taken the course.

For four groups in Table 2 and 2A, either a significant estimate became insignificant or the other way around. But in terms of sign, only edu 'college' had its sign reversal as it has $m \simeq n - m$. The differences between Table 2 and 2A resulted in the ensuing differences between Table 3 and 3A. But the only notable change is that training 'finance/ins.' became significant with its p-value 0.048 in Table 3A under $\Gamma = 1.5$, which is 0.061 in Table 3. Thus, using the D group makes hardly any difference for the first sensitivity analysis. This is because the first sensitivity analysis is "rough". We will see shortly that adding the D group makes more difference for the second sensitivity analysis.

4.2 Second Sensitivity Analysis

Tables 4-1 to 4-9 show the second sensitivity analysis with two parameters $\Gamma = \exp(|\gamma|)$ and $\Delta = \exp(|\delta|)$. For instance, in Table 4-1 for middle school, both the first row for $\Delta = 1$ and the first column for $\Gamma = 1$ have p-value 0.061, which is the same as the no-hidden-bias p-value in Table 3; recall that there is no hidden bias in the second sensitivity analysis when $\gamma\delta = 0$ even if $\gamma \neq 0$, differently from the first sensitivity analysis. Since Tables 4-1 to 4-9 are symmetric about the diagonal, the lower-triangle parts are omitted. Also, the groups with its p-value bound equal to or greater than 0.05 at $\Gamma = 1.5$ or $\Delta = 1.5$ are omitted with the exception of Table 4-1 for middle school. The reason for this exception is that the middle-school group's finding is judged to be insensitive in the third sensitivity analysis below, and thus Table 4-1 is shown for the sake of comparison.

As Γ or Δ increases, the upper bound for the p-value increases. Comparing the tables to Table 3, clearly the first sensitivity analysis is too conservative: e.g., for group 'high school',

$\Gamma = 1.5$ yields NI in Table 3, but 0.000 under $\Delta = 1.5$ and 0.012 under $\Delta = 2.5$ in Table 4-2. Under 5% level, if we adopt the thresholds $\Gamma = 2$ and $\Delta = 2$, among Tables 4-1 to 4-9, only the *three groups' (high school, training 'service', and training 'finance/insurance') effects are insensitive to hidden bias* in the second sensitivity analysis.

TABLE 4-1: Middle School

		Γ				
		1	1.5	2	2.5	3
	1	.061	.061	.061	.061	.061
	1.5		.079	.094	.105	.160
Δ	2			.121	.142	.160
	2.5				.172	.196
	3					.224

TABLE 4-2: High School

		Γ				
		1	1.5	2	2.5	3
	1	.000	.000	.000	.000	.000
	1.5		.000	.004	.012	.026
Δ	2			.042	.092	NI
	2.5				NI	NI
	3					NI

TABLE 4-3: Junior College

		Γ				
		1	1.5	2	2.5	3
	1	.004	.004	.004	.004	.004
	1.5		.020	.047	.074	.097
Δ	2			.116	.153	NI
	2.5				NI	NI
	3					NI

TABLE 4-4: Job 'Clerical'

		Γ				
		1	1.5	2	2.5	3
	1	.000	.000	.000	.000	.000
	1.5		.006	.027	.058	.087
Δ	2			.106	NI	NI
	2.5				NI	NI
	3					NI

TABLE 4-5: Job 'Menial Labor'

		Γ				
		1	1.5	2	2.5	3
	1	.005	.005	.005	.005	.005
	1.5		.017	.032	.048	.062
Δ	2			.075	.115	.144
	2.5				.160	.173
	3					NI

TABLE 4-6: Training 'Info./Comm.'

		Γ				
		1	1.5	2	2.5	3
	1	.002	.002	.002	.002	.002
	1.5		.016	.041	.068	.092
Δ	2			.111	.147	NI
	2.5				NI	NI
	3					NI

TABLE 4-7: Training 'Industrial Appl.'

		Γ				
		1	1.5	2	2.5	3
	1	.007	.007	.007	.007	.007
	1.5		.018	.030	.042	.053
Δ	2			.062	.093	.119
	2.5				.137	.167
	3					.188

TABLE 4-8: Training 'Service'

		Γ				
		1	1.5	2	2.5	3
	1	.000	.000	.000	.000	.000
	1.5		.000	.000	.000	.000
Δ	2			.000	.000	.000
	2.5				.000	.000
	3					.000

TABLE 4-9: Training 'Finance/Insurance'

		Γ				
		1	1.5	2	2.5	3
	1	.006	.006	.006	.006	.006
	1.5		.011	.015	.019	.023
Δ	2			.025	.036	.045
	2.5				.053	.069
	3					.092

The results with the D group included in the C group are in the appendix where Table 4-1A is omitted as it is exactly the same as Table 4-1 and Tables 4-10A and 4-11A are new. Differently from the first sensitivity analysis, the second sensitivity analysis shows more changes by making use of the D group, because the second sensitivity analysis is more "refined" than the first, and because the D group is similar to the T group at the baseline, favoring a D-group woman rather than a C-group woman as a matched control. With $\Delta = \Gamma = 2$ as the thresholds, now the following eight group effects are insensitive: *high school, junior college, job 'clerical', job 'menial labor', job 'mechanic', training 'industrial appl.', training 'service', and training 'finance/insurance'*

4.3 Third Sensitivity Analysis

Table 5 presents the third sensitivity analysis. Since the dimension of x is large, we use the probit propensity score $\Phi(x'a_N)$ instead of x itself for the nonparametric estimation; a similar strategy was used in Lee (2004) where $x'a_N$ was used instead of $\Phi(x'a_N)$. Since $\Phi(x'a_N)$ is one-to-one to $x'a_N$, there is little difference in using $\Phi(x'a_N)$ or $x'a_N$ in the nonparametric estimation; but $\Phi(x'a_N)$ is used here, for it is bounded by $(0,1)$. The bandwidth for each group was chosen by cross-validation plus visual inspection drawing the proportional effect for each group as a function of $\Phi(x'a_N)$. That is, initially cross-validation was done, but if this gave too smooth or too rough a graph, then visual inspection was adopted instead. The bootstrap repetition was set at a relatively low number 200 using only 10% of the C group, because cross-validation and bootstrap nonparametric estimation took way too much time with $N_c = 47060$.

TABLE 5: Proportional Effect and Sensitivity Analysis

	95% LB	estimate	95% UB	G_U^{-1}	G_L^{-1}
edu.: middle school	-0.564	-0.528	-0.106	*1.12*	*2.29*
edu: high school	-0.098	-0.094	0.003	*0.997*	*1.11*
edu.: junior college	-0.135	-0.118	-0.014	*1.01*	*1.16*
edu.: college	-0.052	-0.038	0.083	*0.92*	*1.05*
job: professional	-0.238	-0.100	0.090	*0.92*	*1.31*
job: semi-professional	-0.163	-0.046	0.097	*0.91*	*1.20*
job: clerical	-0.045	0.007	0.061	*0.94*	*1.05*
job: service/sales	-0.109	0.030	0.136	*0.88*	*1.12*
job: mechanic	-0.189	-0.094	-0.010	*1.01*	*1.23*
job: machine operator	-0.412	-0.156	-0.003	*1.00*	*1.70*
job: menial labor	-0.121	0.049	0.101	*0.91*	*1.14*
training: textile	-0.377	-0.180	0.045	*0.96*	*1.61*
training: machine/equip.	-0.025	0.040	0.452	*0.69*	*1.03*
training: info./comm.	0.085	0.175	0.267	*0.79*	*0.92*
training: industrial appl.	-0.189	-0.019	0.130	*0.89*	*1.23*
training: service	-0.284	-0.196	-0.160	*1.19*	*1.40*
training: clerical	-0.037	0.013	0.129	*0.89*	*1.04*
training: finance/ins.	-0.034	0.079	0.504	*0.67*	*1.04*
training: health	-0.118	0.020	0.125	*0.89*	*1.13*

In Table 5, column '95% LB' is the lower 95% confidence bound for the proportional effect, and column '95% UB' is the upper 95% confidence bound; in the notation of Subsection 2.3, 95% LB is $G_L - 1$ and 95% UB is $G_U - 1$. For some small-sample groups in Table 5, the proportional effect estimate in column ‘estimate’ falls near the confidence bounds, partly because the exponential function is asymmetric around zero. The confidence intervals are under selection on observables, and they indicate that six groups have significant proportional effects: middle school, junior college, mechanic, machine operator, info./comm., and service.

The last two columns of Table 5 shows the third sensitivity analysis where the sensitivity parameter is the distance between one and the interval (G_U^{-1}, G_L^{-1}). Among the six groups, (G_U^{-1}, G_L^{-1}) seems too close to one for junior college, mechanic, and machine operator (only 0.01 difference or less). Adopting ξ in (2.4) between 0.01 and 0.08 leads to the conclusion that only *three groups (middle school, training ‘information/communication’, and training ‘service’) have effects insensitive to hidden bias* in the third sensitivity analysis. In the sense that three groups are left, the range $\xi \geq 0.08$ is comparable to $\Gamma = \Delta \geq 2$ in the preceding sensitivity analysis. Using $\xi \geq 0.12$ makes only training ‘service’ effect to be insensitive, which would be too conservative in view of the preceding two sensitivity analyses.

In the appendix, Table 5A repeats Table 5 with the D group included in the C group. Adding the D group to the C group resulted in only small changes for the confidence intervals, although the estimates tend to change more than the confidence intervals. This could be because only 10% of the C group were used in the bootstrap confidence interval construction.

In short, the first sensitivity analysis seems too conservative with only one group (training ‘service’) effect being insensitive to hidden bias; this group’s effect is also insensitive in the other two approaches. In the second and third sensitivity analyses, groups ‘middle school’ and ‘high school’, trainings ‘finance/insurance’ and ‘information/communication’ are insensitive in one approach but not so in the other. This discrepancy between the second and third sensitivity analyses became greater when the D group was added to the C group, which results in five more group effects insensitive to hidden bias according to the second sensitivity analysis: junior college, job ‘clerical’, job ‘menial labor’, job ‘mechanic’, and training ‘industrial appl.’.

5 Conclusions

We conducted three sensitivity analyses for the findings in Lee and Lee (2005, LL) obtained under no hidden bias. Several significant effects in LL lost their significance as hidden bias is allowed for, whereas some still survived the hidden bias. Among the job training types which are of main interest for LL and this paper, 'info./comm.' and 'finance/ins.' trainings seem to have positive effects on re-employment despite hidden bias, whereas 'service' and possibly 'industrial appl.' have negative effects. The three approaches are mostly nonparametric and fairly new, while conventional parametric selection models are still dominant in practice in allowing for hidden bias.

In social sciences, most data are observational, and presence of unobserved confounders is a rule rather than an exception. It is worth emphasizing that unobservables are unobserved, and there is very little one can do about them unless assumptions are imposed. In view of this, the "near-nonparametric" sensitivity analyses are by no means trivial. If more track records for the sensitivity parameters are established in future through more applications so that researchers can agree on how big is big for the sensitivity parameters, then the sensitivity analyses may become useful tools in dealing with unobserved confounders.

REFERENCES

Aakvik, A., 2001, Bounding a matching estimator: the case of a Norwegian training program, Oxford Bulletin of Economics and Statistics 63, 115-143.

Altonji, J.G., T.E. Elder, and C.R. Taber, 2005, Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools, Journal of Political Economy 113, 151-184.

Angrist, J.D. and A.B. Krueger, 1999, Empirical strategies in labor economics, in Handbook of Labor Economics 3A, edited by O.C. Ashenfelter and D. Card, North-Holland.

Caliendo, M., R. Hujer, and S.L. Thomsen, 2005, The employment effects of job creation schemes in Germany: a microeconometric evaluation, IZA Discussion Paper 1512.

Gastwirth, J.L., A.M. Krieger, and P.R. Rosenbaum, 1998, Dual and simultaneous sensitivity analysis for matched pairs, Biometrika 85, 907-920.

Heckman, J.J., H. Ichimura, J. Smith and P.E. Todd, 1998, Characterizing selection bias using experimental data, Econometrica 66, 1017-1098.

Heckman, J.J., R.J. Lalonde, and J.A. Smith, 1999, The economics and econometrics of active labor market programs, in Handbook of Labor Economics 3B, edited by O.C. Ashenfelter and D. Card, North-Holland.

Heckman, J., J.L. Tobias, and E. Vytlacil, 2003, Simple estimators for treatment parameters in a latent-variable framework, Review of Economics and Statistics 85, 748-755.

Hujer, R., M. Caliendo, S.L. Thomsen, 2004, New evidence on the effects of job creation scheme in Germany—a matching approach with threefold heterogeneity, Research in Economics 58, 257-302.

Hujer, R. and S.L. Thomsen, 2006, How do employment effects of job creation schemes differ with respect to the foregoing unemployment duration, ZEW Discussion Paper 06-47.

Ichino, A., F. Mealli, and T. Nannicini, 2007, From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity, Journal of Applied Econometrics, forthcoming.

Imbens, G.W., 2003, Sensitivity to exogeneity assumptions in program evaluation, American Economic Review (Papers and Proceedings) 93, 126-132.

Imbens, G.W. and P.R. Rosenbaum, 2005, Robust, accurate confidence intervals with a weak instrument: quarter of birth and education, Journal of the Royal Statistical Society (Series A) 168, 109-126.

Lee, M.J., 2004, Selection correction and sensitivity analysis for ordered treatment effect on count response, Journal of Applied Econometrics 19, 323-337.

Lee, M.J., 2005, Micro-econometrics for policy, program, and treatment effects, Oxford University Press.

Lee, M.J. and S.J. Lee, 2005, Analysis of job-training effects on Korean women, Journal of Applied Econometrics 20, 549-562.

Manski, C.F., 2003, Partial identification of probability distributions, Springer-Verlag.

Rosenbaum, P.R., 1987, Sensitivity analysis for certain permutation inferences in matched observational studies, Biometrika 74, 13-26.

Rosenbaum, P.R., 2002, Observational studies, 2nd ed., Springer-Verlag.

Rosenbaum, P.R. and D.B. Rubin, 1983a, The central role of the propensity score in observational studies for causal effects, Biometrika 70, 41-55.

Rosenbaum, P.R. and D.B. Rubin, 1983b, 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.

Rubin, D.B., 1974, Estimating causal effects of treatments in randomized and nonrandomized studies, Journal of Educational Psychology 66, 688-701.

APPENDIX

Table 2A: No Hidden Bias Results and Discordant Pairs n (D in C)

variables	est.	t-value	# matched treated	n	m	$n-m$	effect
edu: middle school	-0.104	-1.85	67	15	4	11	-
edu: high school	-0.079	-6.28	2,408	939	374	565	-
edu: junior college	-0.094	-4.11	786	332	129	203	-
edu: college	-0.006	-0.22	726	324	160	164	-
job: professional	-0.059	-1.06	136	58	25	33	-
job: semi-professional	-0.071	-1.66	224	94	39	55	-
job: clerical	-0.066	-4.60	1,997	834	351	483	-
job: service/sales	-0.054	-2.14	557	198	84	114	-
job: mechanic	-0.082	-3.16	548	207	81	126	-
job: machine operator	-0.051	-0.66	59	21	9	12	-
job: menial labor	-0.093	-3.16	471	198	77	121	-
training: textile	-0.146	-2.41	103	41	13	28	-
training: machine/equip.	0.040	0.63	100	40	22	18	+
training: info./comm.	0.049	2.33	962	411	229	182	+
training: industrial appl.	-0.134	-3.26	247	107	37	70	-
training: service	-0.162	-10.7	1,440	525	146	379	-
training: clerical	-0.026	-1.07	687	282	132	150	-
training: finance/ins.	0.209	2.91	67	26	20	6	+
training: health	-0.046	-1.44	392	158	70	88	-

TABLE 3A: Sensitivity for Treatment Assignment (D in C)				
	Γ=1	Γ=1.5	Γ=2	Γ=2.5
edu.: middle school	0.061	0.185	0.284	0.328
edu.: high school	0.000	0.077	NI	NI
edu.: junior college	0.000	0.108	NI	NI
edu.: college	0.434	NI	NI	NI
job: professional	0.180	NI	NI	NI
job: semi-professional	0.061	NI	NI	NI
job: clerical	0.000	NI	NI	NI
job: service/sales	0.020	NI	NI	NI
job: mechanic	0.010	0.144	NI	NI
job: machine operator	0.336	NI	NI	NI
job: menial labor	0.001	0.142	NI	NI
training: textile	0.014	0.126	0.225	NI
training: machine/equip.	0.320	NI	NI	NI
training: info./comm.	0.012	NI	NI	NI
training: industrial appl.	0.001	0.084	NI	NI
training: service	0.000	0.000	0.002	0.058
training: clerical	0.156	NI	NI	NI
training: finance/ins.	0.005	0.048	0.132	0.207
training: health	0.088	NI	NI	NI

TABLE 4-2A: High School (D in C)						
		Γ				
		1	1.5	2	2.5	3
	1	.000	.000	.000	.000	.000
	1.5		.000	.000	.000	.000
Δ	2			.001	.013	.040
	2.5				.064	NI
	3					NI

TABLE 4-3A: Junior College (D in C)						
		Γ				
		1	1.5	2	2.5	3
	1	.000	.000	.000	.000	.000
	1.5		.000	.002	.005	.009
Δ	2			.014	.040	.069
	2.5				.091	.117
	3					NI

TABLE 4-4A: Job 'Clerical' (D in C)

		Γ				
		1	1.5	2	2.5	3
	1	.000	.000	.000	.000	.000
	1.5		.000	.003	.011	.023
Δ	2			.038	.089	NI
	2.5				NI	NI
	3					NI

TABLE 4-5A: Job 'Menial Labor' (D in C)

		Γ				
		1	1.5	2	2.5	3
	1	.001	.001	.001	.001	.001
	1.5		.005	.013	.023	.033
Δ	2			.043	.079	.109
	2.5				.129	.149
	3					NI

TABLE 4-6A: Training 'Info./Comm.' (D in C)

		Γ				
		1	1.5	2	2.5	3
	1	.012	.012	.012	.012	.012
	1.5		.055	.107	.144	.164
Δ	2			.171	NI	NI
	2.5				NI	NI
	3					NI

TABLE 4-7A: Training 'Ind.Appl.' (D in C)

		Γ				
		1	1.5	2	2.5	3
	1	.001	.001	.001	.001	.001
	1.5		.003	.007	.011	.015
Δ	2			.019	.035	.053
	2.5				.068	.098
	3					.134

TABLE 4-8A: Training 'Service' (D in C)

		Γ				
		1	1.5	2	2.5	3
	1	.000	.000	.000	.000	.000
	1.5		.000	.000	.000	.000
Δ	2			.000	.000	.000
	2.5				.000	.000
	3					.000

TABLE 4-9A: Training 'Finance/Ins.' (D in C)

		Γ				
		1	1.5	2	2.5	3
	1	.005	.005	.005	.005	.005
	1.5		.008	.011	.014	.017
Δ	2			.019	.027	.035
	2.5				.041	.054
	3					.073

TABLE 4-10A: Job 'Mechanic' (D in C)

		Γ				
		1	1.5	2	2.5	3
	1	.010	.010	.010	.010	.010
	1.5		.006	.014	.024	.035
Δ	2			.045	.082	.113
	2.5				.132	.149
	3					NI

TABLE 4-11A: Training 'Textile' (D in C)

		Γ				
		1	1.5	2	2.5	3
	1	.014	.014	.014	.014	.014
	1.5		.024	.034	.043	.051
Δ	2			.057	.079	.097
	2.5				.111	.139
	3					.170

TABLE 5A: Proportional Effect and Sensitivity Analysis (D in C)

variables	95% LB	est.	95% UB	G_U^{-1}	G_L^{-1}
edu: middle school	-0.541	-0.338	-0.163	1.20	2.18
edu: high school	-0.109	-0.060	-0.005	1.01	1.12
edu: junior college	-0.126	-0.056	-0.020	1.02	1.14
edu: college	-0.051	0.030	0.070	0.94	1.05
job: professional	-0.219	-0.009	0.052	0.95	1.28
job: semi-professional	-0.176	0.038	0.076	0.93	1.21
job: clerical	-0.048	-0.016	0.041	0.96	1.05
job: service/sales	-0.112	-0.066	0.112	0.90	1.13
job: mechanic	-0.199	-0.090	-0.01	1.01	1.25
job: machine operator	-0.410	0.024	0.008	0.99	1.70
job: menial labor	-0.108	-0.064	0.083	0.92	1.12
training: textile	-0.382	-0.248	0.014	0.99	1.62
training: machine/equip.	-0.052	0.296	0.408	0.71	1.06
training: info./comm.	0.095	0.164	0.261	0.79	0.91
training: industrial appl.	-0.168	-0.058	0.165	0.86	1.20
training: service	-0.281	-0.238	-0.163	1.20	1.39
training: clerical	-0.059	0.018	0.133	0.88	1.06
training: finance/ins.	-0.043	0.373	0.525	0.66	1.05
training: health	-0.104	-0.017	0.107	0.90	1.12

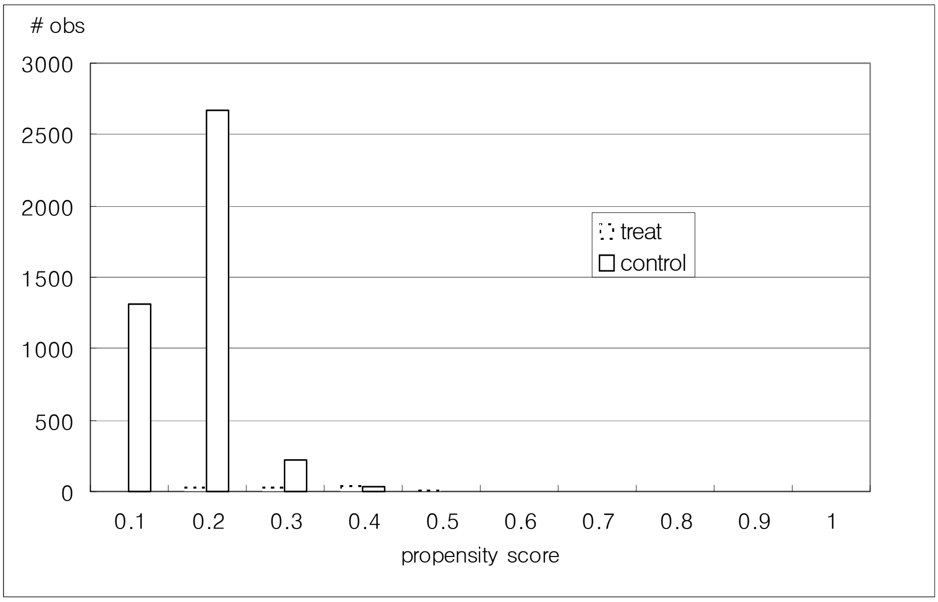

Figure 1: Middle School

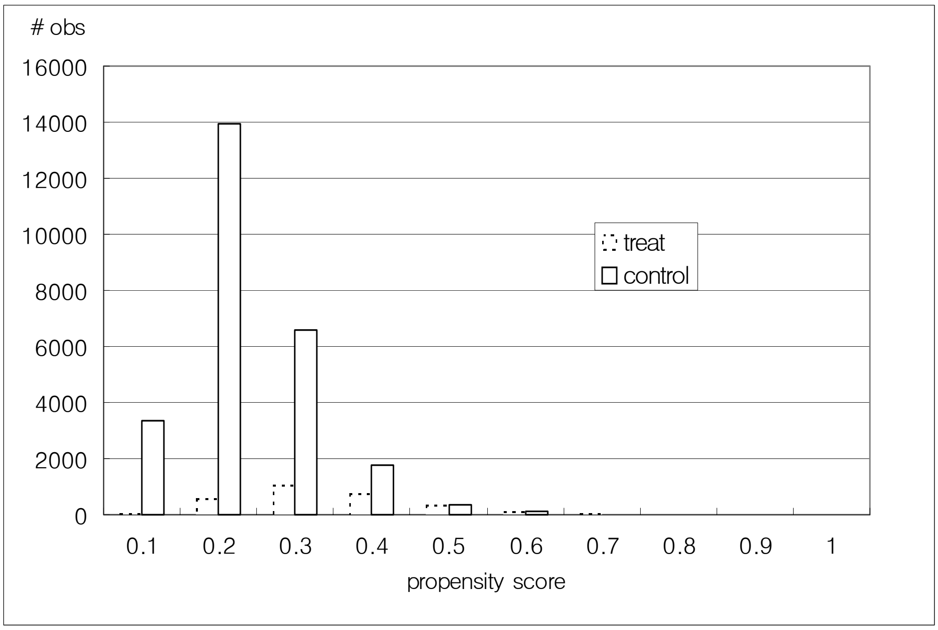

Figure 2: High School

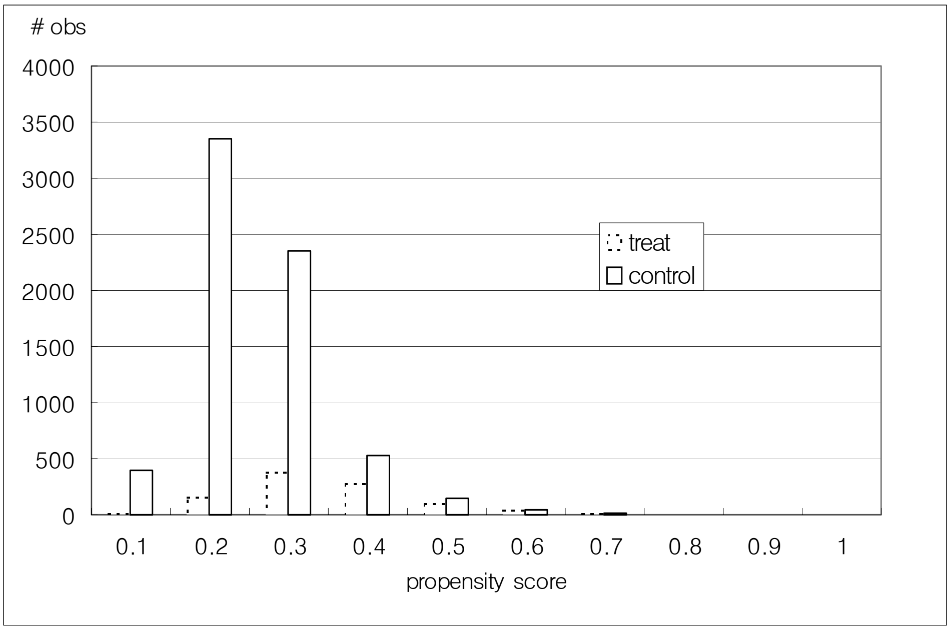

Figure 3: Junior College

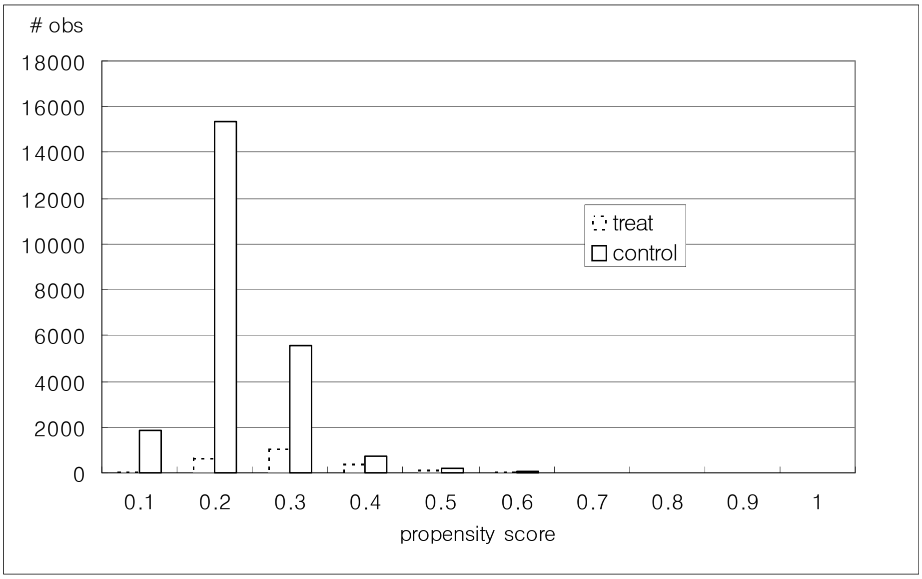

Figure 4: Job: Clerical

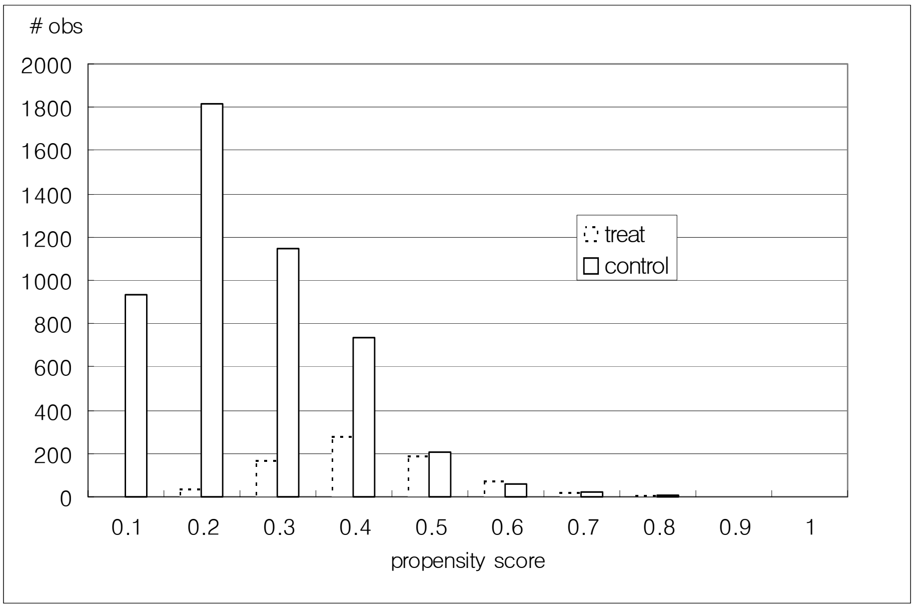

Figure 5: Job: Mechanic

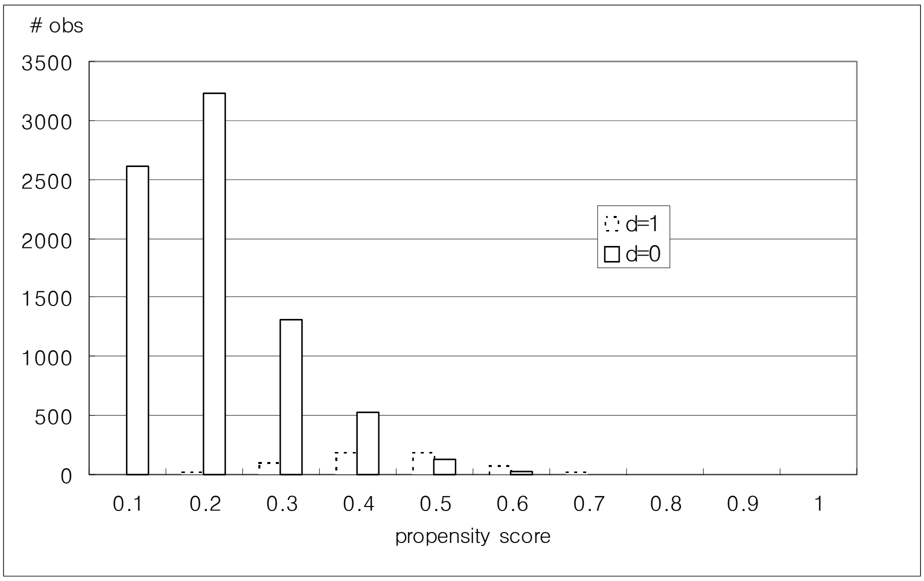

Figure 6: Job: Menial Labor

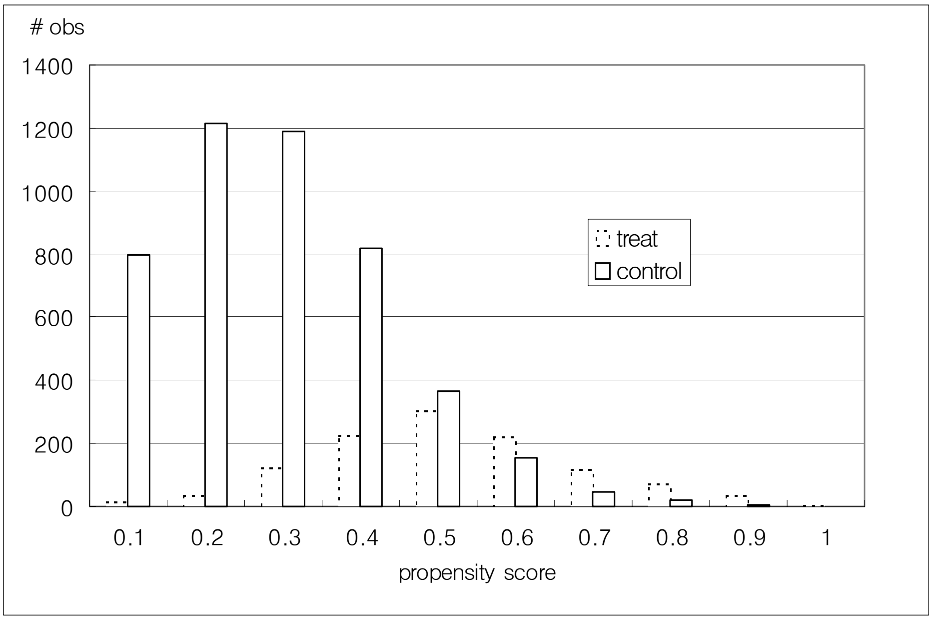

Figure 7: Traning: Info./Comm.

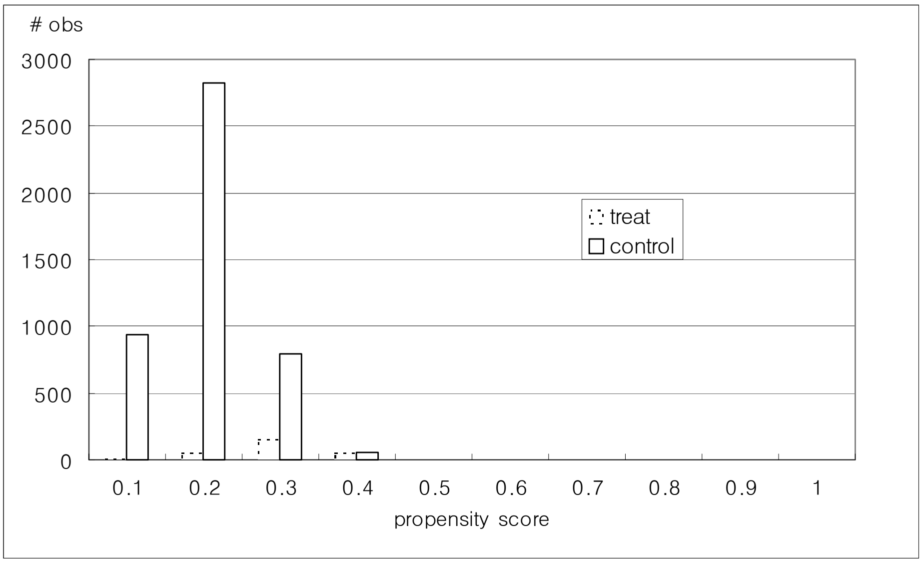

Figure 8: Traning: Industrial Appl.

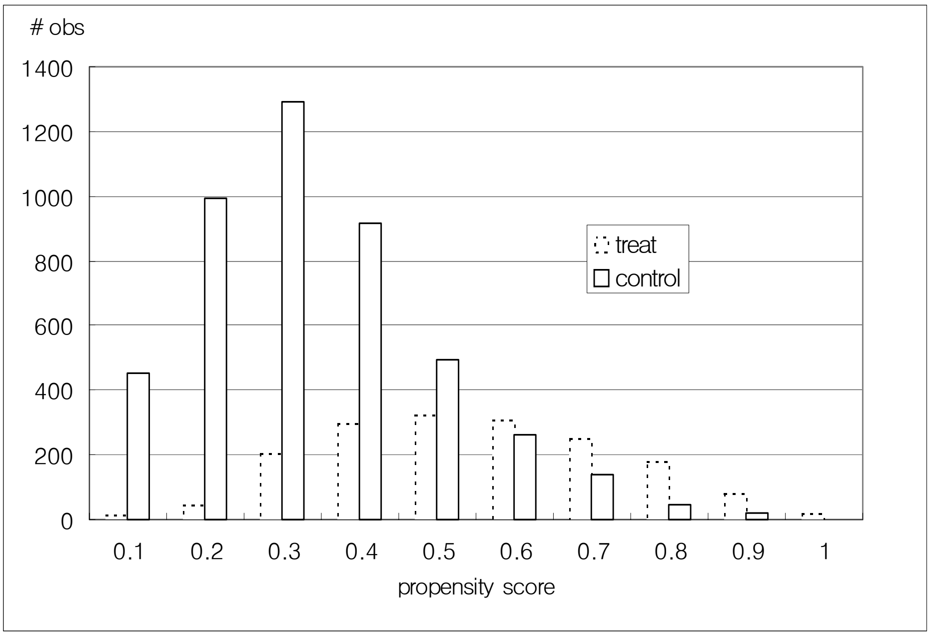

Figure 9: Training: Service

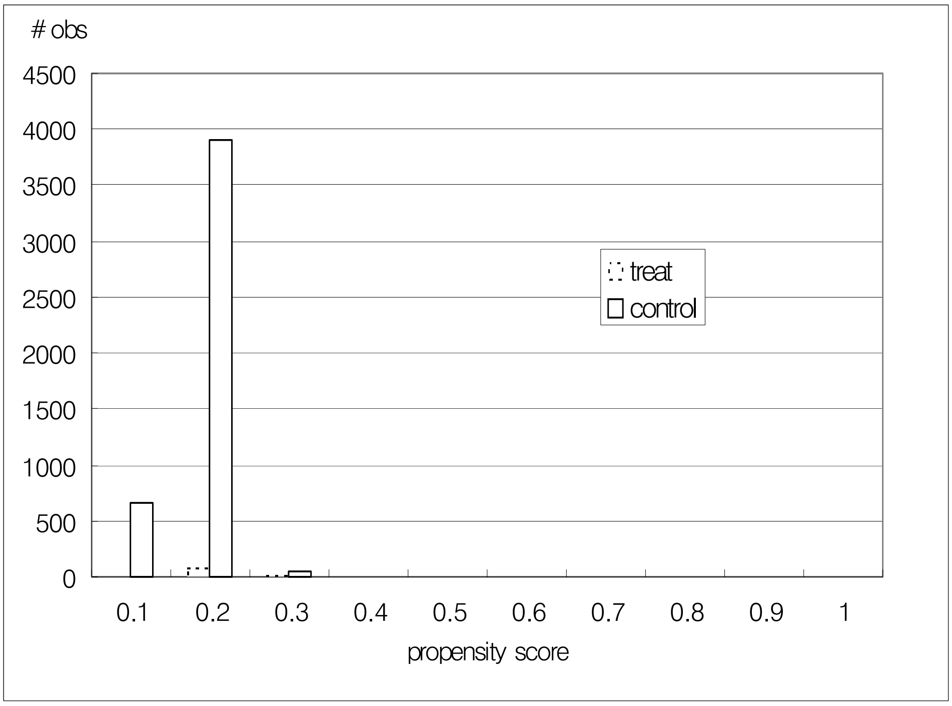

Figure 10: Training: Finance/Ins.