OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 63, 1 (2001) 0305-9049

Bounding a matching estimator: the case of a Norwegian training program

ARILD AAKVIK[*]

University of Bergen

I. Introduction

Interest in training programs both among economists and policy makers remains strong despite many research programs focusing on the effects of training. Among policy makers, the main issue is whether different training programs for unemployed individuals work or not, and for what group it is most effective. Among economists, the discussion has centred around the issue of selection bias and the construction of comparison groups. Assessing the benefits of training programs is difficult since it requires an assessment of what the outcome for trainees would have been without training. Counterfactuals of this sort are never directly observed, and to be able to estimate the training effect some assumptions must inevitably be imposed on the model.

In a classical experiment eligible participants are randomly selected to either a treatment group or a control group. The experiment has the potential of making the treatment and control group equal in every respect. The mean effect of training is found by comparing the mean outcomes of trainees and controls at a given time after training.[1] Non experimental studies differ from randomised experiments in that the probability of participating in a training program is not a fixed constant but influenced by unobserved and observed characteristics due to self-selection and selection made by case workers and

[*] The Norwegian Social Science Data Service in Bergen provided the data, and I would like to thank Dag Kiberg for the work he has done. The data are gathered from sources at The Directorate of Labour, The Social Insurance Oragnaization and Statistics Norway. None of the mentioned institutions are responsible for the analysis or for the conclusions drawn from it.

[1] If success is measured against employment and unemployment spells after training in a random experiment, mean comparisons produce biased training estimates because the sorting into subsequent spells may be different for the treatment group and control group, see for instance (Ham and LaLonde 1996). Non-experimental methods must be used to estimate the training effect in such circumstances.

managers of training programs. A direct comparison of mean outcomes in nonexperimental evaluation studies may either overestimate or underestimate the true effect of training programs. Increasing the sample size is not a remedy for the selection problem.

Selection bias due to correlation between observed variables and a person's training status is solved by either matching techniques or by including these variables in regression analysis. Selection bias due to correlation between unobserved variables and a person's training status is more difficult to remedy. If for instance highly motivated persons are selected for training, and these persons have good labour market prospects regardless of training status, then an observed positive correlation between training status and employment outcomes would not represent a causal effect of training.[2]

In this paper we evaluate a Norwegian vocational rehabilitation program by comparing employment outcomes of trainees and nonparticipants using nonexperimental data. A matching estimator based on the propensity score is used to calculate the training effect for different subgroups of the sample. We demonstrate how bounding a matching estimator can be used to evaluate the intrinsic uncertainty of estimated training effects due to selection on unobservables, using a procedure proposed by (Rosenbaum 1995).

We find that the overall training effect is significant and around six percentage points when we adjust for observed differences between the trainees and the members of the comparison group. The training effect is higher for individuals who are less likely to participate in a training program compared to those who have a relatively high training probability. We also find that employment rates are higher for individuals who are more likely to participate in a training program. This indicates (potentially harmful) 'creaming' and self-selection in the Norwegian vocational rehabilitation (VR) sector.

We scrutinise the estimated training effects to see whether they are sensitive to selection bias due to correlation between unobserved factors and a person's training status. Such an analysis is carried out by calculating upper and lower bounds for different values of unobserved selection bias for the test-statistics under the null hypothesis of no training effect. We find that the overall training effect is sensitive to selection bias. However, the result that

[2] (Heckman 1979) proposed a method that can be used to adjust for correlation between unobserved variables important for the employment and training outcomes, see also (Heckman and Robb 1985). Difference-in-difference models have also been used to correct for selection bias, see for instance (Ashenfelter 1978), (Bassi 1983), (Ashenfelter and Card 1985), (Card and Sullivan 1988), and (Heckman and Hotz 1989). Recently, different nonparametric and semiparametric specifications of the Heckman two-equation selectivity adjustment technique have been proposed, see for instance (Heckman and Honoré 1990), (Newey, Powell and Walker 1990), (Ahn and Powell 1993), and (Heckman, Ichimura, Smith and Todd 1998).

the training effect is significant and positive for individuals who are less likely to participate in a training program is not sensitive selection bias. These results have important policy implications for the VR sector in Norway: There is scope for improving the efficiency (and redistribution effect) of training by selecting the most 'needed', since these individuals are more likely to reap the benefit of training, although they have lower employment rates without training.

In the next section we discuss the data used in this paper. In section III we discuss selection bias, and set up a framework of which training effects can be estimated. We start by assuming that selection on unobservables is ignorable, that is, we assume that selection bias is only due to correlation between observed background variables and a person's training status. We check for nonoverlapping regions of the propensity score for trainees and nonparticipants, and whether the distributions of individual background characteristics differ for trainees and nonparticipants, see (Heckman *et al.* 1998). These are the main sources of selection bias due to observed variables. We estimate the training effect for different strata where strata are constructed by pairing participants and nonparticipants with approximately common probabilities of training. In section IV we discuss selection bias due to unobserved variables. We scrutinize the estimated training effects to see if they are sensitive to selection bias due to correlation between unobserved factors and a person's training status. This is done by bounding the matching estimator. Section V summarizes the results.

II The Data and Institutional Settings

The Norwegian vocational rehabilitation sector offers income maintenance payments and training programs for individuals with reduced productivity in the labour market due to medical conditions. The VR sector has expanded rapidly since the national social insurance act was passed in 1966. The expansion has neither been guided by a firm knowledge on the overall economic impact of the training programs, nor on which groups may benefit most from program participation. Today, around 35,000 persons participate in a VR training program each day, which is around 1.5 percent of the labour force.

Most persons who apply for a VR job training program have previously been employed. Sickness benefit is usually given while the person is still employed in the old job. Medical treatment is given during this period. The sickness benefit in Norway is generous, paying 100 percent of previous income for up to 52 weeks, subject to a maximum benefit restriction of around NOK 235,000 (USD 33,000). State and municipal employees and many employees in large companies have collective agreements stipulating that the employer is to make up the difference between the employee's wage and sickness benefit. This secures that even high income earners receive full pay during illness.

Individuals unable to return to work after 52 weeks on sickness benefits are entitled to a VR benefit. The decision to accept VR benefit is made by the local Social Security Office, usually after a recommendation from a medical doctor.[3] The VR benefit is usually two-thirds of the gross income in the previous year subject to maximum and minimum benefit restrictions, and is thus less generous than the sickness benefit. Health status is the legal eligibility criterion for VR benefit, but labour market prospects and social integration may also implicitly be taken into account by the local Social Security Office or the medical doctor. Waiting periods exist neither for episodes between work and sickness benefit nor between sickness benefit and VR benefit. In 1989 there was no maximum number of weeks on VR benefit but normally episodes do not exceed 3–4 years.

While receiving VR benefit, a decision has to be made whether the individual can return to the old job or has to search for a new job. At this stage, some people return to their old job or apply for a disability pension without entering the training sector. Individuals who are not granted a disability pension or by their own effort return to their old job, are referred to the local Employment Office for participation in a job training program. This referral and application process is usually assisted by the local Social Security Office and medical doctors. The local Employment office evaluates whether training may help applicants obtain a job. The process at the local Employment office starts with a conversation between a VR labour consultant and the VR client. The consultant asks about interests and potential occupations, and the severity of medical conditions.[4] The case worker and each client usually decides on a rehabilitation plan which includes participation in one or more training programs, where the final goal is to place the client into a new job.

The decision to accept a person into a training program is mainly taken by case workers at the Employment Office and local managers of vocational rehabilitation centres. This decision is usually based on subjective judgement regarding employment prospects. In a report written by The National Social

[3] After our observation period ended in 1993 there where made some institutional changes in VR responsibilities. As from 1994, the labour market authorities decide both on rehabilitation benefit payments and training participation.

[4] Each individual has at least one medical diagnosis, for instance 'hardness of hearing', 'lower back injuries', 'migraine', 'alcoholism', 'drug abuse', 'minor mental disorders', 'problems in social adjustment', etc. There is a clear administrative distinction between medical rehabilitiation (MR) and vocational rehabilitation (VR) in Norway. Even though health improvements may occur during vocational rehabilitation, the main purpose of VR training programs is to enhance employability given medical diagnosis, not to improve health impairments.

Insurance Organization, (RTV 1985), it was emphasized that an 'evaluation of the clients' total situation in each case should be considered when a participation decision is made. Main inclusion criteria are health, age, personal characteristics, social conditions, education and labour force experience.' Managers of training programs interviewed by (Ford 1993) believed that rehabilitation efforts aimed at elderly clients (above the age of 50) are of scant value in a labour market with increasing unemployment. Our data confirm that younger clients have a higher probability of training participation.[5] The candidates for training participation may also influence the participation by subjective information supplied to the program administrator and case workers. A third possibility is that eligible persons may be assigned to program participation on a roughly first-come, first-serve basis. Such an enrolment rule would resemble a controlled experiment, where the training effect is estimated as the mean difference in outcome for individuals randomly selected for treatment and control groups.

The training programs offered are typically ordinary education (classroom training), work training in factories, or more traditional vocational training courses sometimes leading up to a job related certificate. Training programs also include wage subsidies and employment in the public sector. The service varies in substance and duration across clients, reflecting a diverse clientele and broad orientation of vocational rehabilitation training programs. All schooling and labour market training are free of charge for the participants, and participants usually receive VR benefit while undergoing training. Program participants may also rely on other benefits, such as unemployment benefit, sickness benefit or social assistance. The benefits cease upon the return to work.

In our analysis we use data on 4416 VR clients that had been referred to the local Employment Office in 1989 for evaluation and training participation. We have relatively detailed information on socioeconomic background, labour market participation, and health status[6] for the persons in our sample. We can observe which clients of those who applied participated in a vocational rehabilitation training program, and who were employed by the end of the observation period.[7]

[5] The unemployment rate increased during our observation period from 1989 until the end of 1993. The unemployment rate reached its peak in late 1993. From then on the unemployment situation improved considerably.

[6] Health status is a variable that consists of 16 different groups of diagnoses. In all the regressions done in our analyses we have included the health status variable, but we do not report in the tables. Descriptive statistics and regressions where we report coefficients on the diagnosis are available on request from the author.

[7] A person is defined as a trainee if (s)he applied for a training program in 1989 and was registered in a training program for at least 5 days in the period from the beginning of 1989 until the end of 1993. A person is a nonparticipant if (s)he applied for a training program in 1989, but is never registered as a participant in a training program. For arguments in favor of 'internal' comparison groups, see (Bell, Orr, Blomquist and Cain 1995).

Successful rehabilitation in the bureaucratic sense does not necessarily mean that the client got a job, but that the client is prepared for ordinary work. In this case, the client becomes an ordinary job seeker eligible for standard service from the local Employment Office. However, in our empirical analysis we use employment by the end of 1993 in an ordinary job as the success criterion.[8]

Before proceeding, we will examine the characteristics of participants and nonparticipants. Table 1 contains descriptive statistics of central variables in the empirical analysis for program participants and comparison group members.

The mean income before the clients enter VR is virtually indentical among the groups. Mean income of spouses is almost 14 percent higher for the comparison group members. The frequency of marriage is also higher among nonparticipants. The average age of participants is two and a half years lower than that of nonparticipants. Participants are also better educated, but only by a small margin. However, participants have less working experience than nonparticipants. Furthermore, participants are more likely to hold a job in 1993. A more formal analysis of the selection process can be conducted within the framework of a logit model. Table 2 reports the results from the logit model.

It is clear from the Chi-squared test that the selection model is significant as measured against the same model, with no explanatory variables. Thus, individuals participating in a program differ significantly from eligible nonparticipants with respect to observable characteristics. We thus reject the first-come, first-serve hypothesis, which would have indicated an insignificant Chi-square statistics.

Notice from the table that several variables are significant in the selection regression. Favourable characteristics like youth, higher education in years, and higher income before VR started, increase significantly the probability of participating in a training program. Higher unemployment rates in local districts also make it more likely that a person participates in a training program. It is expected that younger individuals with more education and higher previous income ex ante are more easily rehabilitated than older less educated individuals with fewer years of job experience. This is consistent with a hypothesis of 'creaming', and also with the information on actual behaviour reported by Ford (1993).

[8] Employment is defined in this paper as working at least 20 hours per week at the end of our observation period, which is 1993. We have experimented with several different definitions of employment outcomes. For instance, we have used full time employed and conditioned on minimum spells of 60 and 90 days in a job in our definition of employment. However, the empirical results are not sensitive to our definition of employment.

TABLE 1
Mean Characteristics

	Participants	*Comparison group*
Number	2908	1508
Employment ratios 1993, percent	38	30
Male, percent	57	55
Married, percent	33	40
Income in 1988 in NOK (a)	72928	72178
Income in 1992 in NOK (a)	69969	60020
Spouse's income in 1988 in NOK (a)	45591	54054
Spouse's income in 1992 in NOK (a)	55094	62784
Parent, percent (b)	42	42
Age of the youngest child (c)	34	33
Unemployment in local districts, 1989 (d)	3.5	3.4
Unemployment in local district, 1992	4.2	4.2
Age	33.5	36.0
Work experience in years (e)	11.4	12.2
Education in years	10.4	10.2

Notes:

(a) Individuals having zero income or spousal income are included in these numbers. NOK is Norwegian Kroners and around 7 NOK buy 1 USD.

(b) This is the dummy variable which has the value one if the person has a child under the age of 11, and zero otherwise. The numbers reported are in percent.

(c) This is a dummy variable which has the value one if the person has a child over the age of three, and zero otherwise. The numbers reported are in percent.

(d) We calculate this unemployment rate as the total number of unemployed individuals multiplied by 100 and divided by the total number of individuals in the age interval between 16 and 67 in the municipality (kommune). In official statistics the number of unemployed individuals is divided by the number of individuals in the labour force, producing a higher rate than ours. The numbers are presented in percent.

(e) Work experience is number of years a person has a yearly income of more than a given level (G) which for instance in 1989 was NOK 32,700 (USD 4,600). G is adjusted for inflation each year, and is the level where you start earning points for an old age pension in the Norwegian Social Insurance scheme.

The presence of children and whether the age of the youngest child is below or above three are also significant inclusion variables in the selection model. There is a marked difference in Norway in the probability of getting day-care placement for children below and over the age of three. Childless individuals experience better chances of being a participant. Everything else being equal, having a child below the age of three decreases the participation probability significantly, which is related to the fact that there is a marked difference in the probability of getting day-care for children below and over the age of three.

TABLE 2
Selection into Training. Logit Model. (a)

	Selection *(1)*	*Marginal effects* *(2)*
Constant	0.568*	0.147*
	(0.281)	(0.063)
Age	−0.020**	−0.004**
	(0.005)	(0.001)
Married (b)	−0.180*	−0.040*
	(0.096)	(0.021)
Male	−0.007	−0.002
	(0.077)	(0.017)
Education (c)	0.047**	0.011**
	(0.018)	(0.004)
Income (d)	0.104*	0.023*
	(0.056)	(0.013)
Income, spouse (d)	0.012	0.003
	(0.052)	(0.012)
Children (e)	−0.290*	−0.065*
	(0.124)	(0.028)
Child's age (f)	0.432**	0.097**
	(0.126)	(0.028)
Unemployment	0.039*	0.009*
	(0.021)	(0.005)
Work experience (g)	0.006	0.001
	(0.008)	(0.002)
Mean probability	67%	
Log-Likelihood	−2788	
Chi-squared	92	

Note: Numbers in parentheses are standard deviations. *Significant at the 10 percent level. **Indicates significance at the 1 percent level, both for a two-sided test of population coefficients equal to zero.

(a) The dummies for diagnosis are included in the regression but not reported here to save space.
(b) This is a dummy variable for whether the person is married or not.
(c) Education is measured in number of years. Regressions where we use other measures of education are available from the author on request.
(d) Income and spouse's income are measured in NOK 100,000 and are measured for 1988.
(e) This is a dummy variable for the presence of children under the age of 11.
(f) This is a dummy variable which has the value one if the person has a child over the age of three and zero otherwise.
(g) Work experience is actual years a person has income larger than NOK 32,700 in 1989.

III Selection on Observed Variables

The model

Let there be n individuals eligible for training. $D_i = 1$ if the ith person receives training, and $D_i = 0$ if this unit receives the control. Fix the n_1

number of individuals in training, where $n_1 = \sum_{i=1}^{n} D_i$. Let D be a random $n \times 1$ matrix containing the D_i for all units, and let Ω be the set that contains all the possible treatment assignments D. n_0 is the number of nonparticipants, and $n = n_1 + n_0$. Furthermore, let $|\Omega|$ be the number of elements in the outcome set Ω, where $|\Omega| = \binom{n}{n_1}$. Each element in the sample space consists of a n-tuple of n_1 ones and $n_0 = n - n_1$ zeros. The outcome of the selection process is the $n \times 1$ matrix d, where $d \in \Omega$.

In a randomized experiment each possible $d \in \Omega$ is given the same probability, that is, $Pr(D = d) = \frac{1}{\Omega}$. If $Pr(D = d) \neq \frac{1}{\Omega}$ for each $d \in \Omega$, then this particular way of determining the assignment of treatments to eligible persons is called a nonrandomized experiment or nonexperiemental data.

If the selection of candidates into training programs is based on observed variables in the data, x, each possible $d \in \Omega$ is given the same conditional probability. All the test statistics developed under the assumption of random sampling are valid given that we in some way can take into account selection based on x, either by matching techniques or by including the variables in regression analysis.

Assume that the training status of the ith person depends only on the treatment assigned to this person, and is unaffected by the particular assignment of treatments to other individuals. This is the assumption of independent training outcomes. (Rubin 1986) calls it SUTVA, for the 'stable unit treatment value assumption'. We also assume independent employment outcomes and thus ignore potential general equilibrium effects of large scale training programs. The realized employment outcome is

$$y_i = d_i y_{1i} + (1 - d_i) y_{0i} \tag{1}$$

where the subscript 1 denotes trainees and 0 members of the comparison group. This model is called the Roy model ((Roy 1951)), or a switching model, see (Quandt 1988). It is also a potential outcome model since we have two potential outcome states; the treated state and the untreated state, and only one observed outcome.

Let the response some time after the treatment be the N-dimensional column vector y consisting of observed outcomes. If the observed response vector y is unaffected by different treatment assignments $d \in \Omega$, the treatment is said to have no effect. If y is different for different $d \in \Omega$, then the treatment has at least some positive (or negative) effect. The estimated treatment effect is said to be significant if a test-statistic $t(d, y)$ crosses some given critical value.

To use a test-statistic like $t(d, y)$, we first have to make the treatment group and the comparison group as equal as possible since the test is based on random sampling. Adjustment for observed selection factors can be carried out within the framework of regression analysis, or by different matching techniques. In this paper we use a matching technique based on the propensity score rather than regression analysis. Individuals are placed into twelve subclasses based on the estimated propensity score, where each stratum consists of 368 individuals. Within each stratum participants and nonparticipants have approximately the same estimated probabilty of training. Since $|\Omega|$ is extremely large, we follow the usual procedure of approximating the distribution of $t(d, y)$ by its mean and variance.

We use the (Mantel and Haenszel 1959) statistics to test the null hypothesis of no training effect. The Mantel-Haenszel nonparametric test can be used to test for no training effect both within each stratum and as a weighted average between strata. The test compares the number of persons in the training group who are employed against the same expected number given that the training effect is zero. Under the null-hypothesis the distribution of y is hypergeometric. Let n_{1s} and n_{0s} be the number of individuals treated and not treated in stratum s, respectively, where $n_s = n_{1s} + n_{0s}$. Let y_{1s} be the number of employed participants, y_{0s} the number of employed nonparticipants, and y_s the total number of employed individuals in stratum s. The test-statistic $Q_{M-H} = (y_{1s} - E(y_{1s})/Var(s_{1y}))$ follows the chi-square distribution with one degree of freedom. The Mantel-Haenszel test-statistics for the overall effect of training are given by

$$Q_{M-H} = \frac{U^2}{Var(U)} \tag{2}$$

where

$$U^2 = \left[\sum_{s=1}^{S}\left(y_{1s} - \frac{n_{1s}y_s}{n_s}\right)\right]^2$$

and

$$Var(U) = \sum_{s=1}^{S}\left(\frac{n_{1s}n_{0s}y_s(n_s - y_s)}{n_s^2(n_s - 1)}\right)$$

which is chi-square distributed with one degree of freedom.

The theory on matching suggests that within each group the background variables are approximately equal for participants and nonparticipants.[9] The propensity score, defined as $\pi_i(x_i) = Pr(D_i = 1|x_i)$, is the conditional probability that an individual with $X = x$ will participate in a training programme. We have calculated the propensity score based on the logit model

[9] (Cochran 1968) shows that five subclasses are often sufficient to remove more than 90 percent of the bias. See also (Rosenbaum and Rubin 1984). Our choice of twelve subclasses is ad hoc; we want to be certain that most of the bias is removed, but at the same time have a sufficient number of participants and nonparticipants in each group.

reported in Table 2. The propensity score is then used to place individuals into different subclasses. Individuals with the lowest estimated propensity scores are placed in group number one, while those 368 persons most likely to participate in a training program are placed in group number twelve.

For almost every participant in a training program, there is a comparable nonparticipant: The group with the highest estimated training probabilities contains 282 persons on training (9.7% of the total numbers of trainees), and 82 persons not in training (5.4% of all nonparticipants), see Figure 1 in the appendix. The group with the lowest estimated probabilities of training consist of 189 people on training (6.5% of all participants), and 179 persons not in training (11.9%). Figure 1 also shows that trainees are increasingly over represented in groups with higher training probabilities, and that nonparticipants are increasingly over represented in groups with lower probabilities. This is not surprising given that the logit model to some extent is successful in explaining the observed selection into training programs. In experimental studies each individual has the same probability of participating in a training program, which would indicate two horizontal lines at 8.33 percent in Figure 1. (Figures 1–17 appear in the Appendix.)

Due to our large number of observations, it is expected that the matching procedure will distribute the mean value of x equally between trainees and participants within each stratum, so the estimated training effect has no bias due to $\bar{x}$. Figures 2 and 3 display the balance within the different strata for the two initially most unbalanced variables in the data set, namely, age and educational level. These two variables had a large and significant impact on the probability of training in the logit model. Differences in mean values of the background variables for trainees and nonparticipants within each stratum are small after matching, not only for the two variables shown in Figures 2 and 3, but also for all other background variables.[10] The two-sample t statistics for comparing training and nontraining group means after matching were not significantly different from zero for all variables and strata.[11]

Even if the matching technique is successful in equating the mean characteristics in the treatment and control groups, it does not necessarily equate

[10] We have also plotted the mean values for all the other variables in our data. There are some differences in the mean values of income and spouse's income for participants and nonparticipants. However, the mean differences are not significantly different from zero within each stratum. The figures are available on request from the author.

[11] There is a trade-off between the difference in the absolute value of the mean for participants and nonparticipants, and its statistically significance. Increasing the number of strata may increase the mean difference for background variables, but make it less likely that they are statistically significantly different from zero, since the number of observations within strata decreases for increasing numbers of strata. However, we do not consider this to be a problem in this paper since we have 368 observations in each stratum.

the distribution of characteristics in the two groups, however, with correct weighting it should. We have plotted the value of the age variable for the treatment and comparison group, and compared the empirical distributions. Figure 4 shows the (unconditional) age distribution for trainees and nonparticipants for the entire sample.[12] Pariticipants are clearly over represented in the lower age groups, and under represented in the higher age groups. Figure 4 also shows that older people are less likely to participate in a training program. Figures 5 to 16 show the within strata age distribution for all the strata constructed in this paper. Group 1 contains people with the lowest probabilities of participating in a training program, while strata 12 contains people with the highest training probabilities. These two stratum are expected to produce the most unequal distribution of individual characteristics for trainees and nonparticipants. In addition, age is the variable for which participants and nonparticipants initially differed the most.

The age distribution for participants and nonparticipants is surprisingly equal for all the strata, except stratum 2, and possibly also stratum 5.[13] Within stratum 2 participants are clearly over represented in the younger age cohorts. The two groups have at least partially unequal distributions within strata 2. However, the age distributions for trainees and comparisons in stratum number 2 and 5 are atypical in our data. The discrepancies in the distribution of different individual characteristics are generally small. The bias that can be attributed to differences in the shape of the distribution of the regressors is largely eliminated by the matching technique in our data.

A third point about the treatment and comparison group used in this paper relates to whether there are differences in the support of the regressors between trainees and comparisons. See (Heckman *et al.* 1998) for a discussion. The matching estimator cannot identify the training effect outside the region on common support. Define $S_{1x} = \{x|f(x|D = 1) > 0\}$ to be the support of x for trainees, where $f(x|D = 1)$ is the density of the background vector conditional on participation in a training program. Let $S_{0x} = \{x|f(x|D = 0) > 0\}$ be the support of x for members of the comparison group. Let S_x be the region of common support where $S_x = S_{1s} \cap S_{0s}$. Similarly, we can define the support of the propensity score for participants of training programs as $S_{1\pi(x)} = \{\pi(x)|f(\pi(x)|D = 1) > 0\}$, and for nonparticipants as $S_{0\pi(x)} = \{\pi(x)|f(x)|D = 0) > 0\}$. The common support region for the propensity score is $S_{\pi(x)} = S_{1\pi(x)} \cap S_{0\pi(x)}$.

The propensity score in our data is defined in the interval (0.38, 0.82), so the propensity is not defined for the whole probability interval between 0 and

[12] We have grouped the age variable into five-year intervals.

[13] More figures for other variables are available on request from the author. These variables show less differences in the shape of the distribution for trainees and participants for different values of the propensity score than age.

1. The support for trainees is $S_{1\pi(x)} = (0.43, 0.82)$, and the support for comparisons is $S_{0\pi(x)} = (0.38, 0.81)$.

Define $S_{1\pi(x)}\backslash S_{\pi(x)}$ to be the support of $\pi(x)$ given $D = 1$ which is not in the overlap set $S_{\pi(x)}$, and $S_{0\pi(x)}\backslash S_{\pi(x)}$ to be the support of $\pi(x)$ given $D = 0$ which is not in the overlap set $S_{\pi(x)}$. Only 3 out of 1508 nonparticipants are outside the common support region of $\pi(x)$, and only 2 out of 2508 participants are outside the common support region of $\pi(x)$. Bias due to nonoverlappping support for $\pi(x)$ is found to be a large part of selection bias in (Heckman *et al.* 1998). In our study, however, selection bias from using the total sample rather than restricting it to regions of common support for the propensity score is negligible.

The discussion above suggests that our data are well suited to adjust for observed selection. The problem of selection bias that can be attributed to comparing incomparable persons, and selection bias that arises from differences in the shape of the distribution of the regressors between the trainees and comparisons are ignorable in our data. (Heckman *et al.* 1998) find that these two components are the most important parts of the conventional measure of selection bias in US data. They find that unobserved selection (selection rigorously defined) is a relatively small and insignificant fraction of selection bias, only 7 percent. However, the matching technique will not eliminate selection bias due to correlation between training and unobserved characteristics. Even though unobserved selection bias may be a small part of selection bias, it may still be an important part compared to the estimated training effect in nonexperimental studies. We will discuss unobserved selection bias in more detail in section IV.

Results

The effectiveness of training can be estimated by comparing employment rates for trainees and nonparticipants within each stratum. This is the effect given that we have taken into account selection bias due to correlation between fifteen observed variables and the training variable. The average training effect within each stratum is calculated as the difference in mean outcomes for the treatment group and comparison groups. Table 3 shows the mean employment rates for participants of training programs and members of the comparison group within each stratum, together with the training effects. The Mantel-Haenszel test statistics is used to test the null hypothesis of no training effect.

The overall effectiveness of training programmes in raising employment among participants is positive and significantly different from zero at the 1 percent level. After matching on the propensity score, and thus adjusting for observed characteristics in the data, we find that the average training effect is

TABLE 3
Estimating the effect of Training on Employment for Sub-samples

Group (a)	*Number*	*Employment rates*	*Difference in empl. rates*	*Q(M−H)*	*Prob. Value*
1 Treatments	189	0.3545			
Comparisons	179	0.1620	+0.1925	17.62	0.000
2 Treatments	206	0.3786			
Comparisons	162	0.2346	+0.1441	8.70	0.003
3 Treatments	225	0.4222			
Comparisons	143	0.2308	+0.1915	14.09	0.000
4 Treatments	220	0.3500			
Comparisons	148	0.3311	+0.0189	0.14	0.708
5 Treatments	247	0.4251			
Comparisons	121	0.2149	+0.2102	15.61	0.000
6 Treatments	259	0.3707			
Comparisons	109	0.2752	+0.0954	3.09	0.078
7 Treatments	246	0.3333			
Comparisons	122	0.3689	−0.0355	0.45	0.500
8 Treatments	257	0.3852			
Comparisons	111	0.3333	+0.0519	0.89	0.345
9 Treatments	254	0.3425			
Comparisons	114	0.2982	+0.0443	0.70	0.403
10 Treatments	257	0.3774			
Comparisons	111	0.4144	−0.0370	0.45	0.505
11 Treatments	266	0.3684			
Comparisons	102	0.4118	−0.0433	0.59	0.444
12 Treatments	282	0.4433			
Comparisons	86	0.5233	−0.0800	1.69	0.193
Total:					
Treatments	2908	0.3793			
Comparisons	1508	0.3165	+0.0628	20.35	0.000

(a) Subclasses are constructed using the estimated propensity score from the logit model reported in Table 2.

6.3 percentage points. This estimate is almost identical to that found in (Aakvik and Risa 1998) using regression analysis, and is almost 2 percentage points lower than the difference in unadjusted employment rates between trainees and comparisons.

If we take a closer look at the estimated average training effect within strata, we find some interesting results. The training effect is positive and significant at the 1 percent level for strata 1, 2, 3 and 5, and as high as 15−20 percentage points. All the other subclasses produce a training effect not different from zero at the 1 percent level, although the training effect in stratum number 6 is positive and significant at the 10 percent level.

The highest employment rates without training are found in the strata with high values of the probability of program participation. However, the training programs are more efficient in raising employment rates for those groups with relatively low training probabilities. The selection process indicated by the observed data, which is a combination of self-selection and selection by case workers, shows that those who have the best chances of being employed independent of training status are over represented in training, while those who benefit most, even though they have lower employment rates without training, are under represented on training programs. This is consistent with 'harmful' cream skimming on observed variables.

IV. Selection on Unobserved Variables

The model in Section III assumed that trainees and comparisons are different because they differ on observed variables in the data set. This is equivalent to assuming that $(y_1, y_0) \perp\!\!\!\perp d|x$, where '$\perp\!\!\!\perp$' denotes independence. If trainees and comparisons differ on unobserved measures, a positive association between a person's training status and employment outcome would not necessarily represent a causal effect. Although we have many background variables, training effects in nonexperimental studies may be contaminated with selection bias due to unobserved factors like motivation, ability, preferences, etc. Given that we have adjusted for selection bias due to nonoverlapping support and discrepancies in the distribution between participants and nonparticipants, the purpose of sensitivity analysis is to ask whether inferences about training effects may be altered by factors not observed in the data. It is not possible to estimate the magnitude of selection bias with nonexperimental data. We rather calculate upper and lower bounds on the test-statistics used to test the null hypothesis of no training effect for different values of unobserved selection bias, following a procedure proposed by (Rosenbaum 1995).

The model

Assume we have the following model to describe the probability of training

$$\pi_i = Pr(D_i = 1|x_i) = F(\beta x_i + \gamma u_i) \tag{3}$$

where x_i is the observed vector of background variables for individual i, u_i is an unobserved variable, and γ is the effect of u_i on the probability of participating in a training programme. If we assume that F is the logistics distribution, the odds that individual i is a participant can be written

$$\left(\frac{\pi_i}{1-\pi_i}\right) = exp(\beta x_i + \gamma u_i) \tag{4}$$

If we compare two persons within a stratum, with common support of x and equal distribution of x, the odds ratio (relative odds) of receiving the treatment may be written

$$\frac{\left(\dfrac{\pi_i}{1-\pi_i}\right)}{\left(\dfrac{\pi_j}{1-\pi_j}\right)} = \frac{\pi_i(1-\pi_j)}{\pi_j(1-\pi_i)} = \frac{exp(\beta x_j + \gamma u_j)}{exp(\beta x_i + \gamma u_i)} = exp[\gamma(u_i - u_j)] \tag{5}$$

where i and j are two different individuals within a stratum. The x-vector cancels since it is approximately equally distributed for all individuals within each stratum. Equation (5) says that if two individuals have different values of u, the difference in the odds of participating involves the parameter γ and the difference in u. If there are no differences in unobserved variables, or if unobserved variables do not influence the probability of participating, the odds ratio is one, which implies no unobserved selection bias. In that case, controlling for observed selection would produce unbiased estimates of the training effect. In sensitivity analysis we evaluate how inference about the training effect will be altered by changing the values of γ and $(u_i - u_j)$.

It is assumed for simplicity, and to get easy interpretable numbers, that the unobserved variable is a dummy variable u with coordinates $u_{si} = 1$ or $u_{si} = 0$ for each (s, i). Suppose that motivation is an important unobserved and omitted factor for both the training decision and employment outcome. In that case the model assumes that a person is either motivated ($u = 1$) or not motivated ($u = 0$). We can rewrite equation (5) into

$$\frac{1}{e^{\gamma}} \leqslant \frac{\pi_i(1-\pi_j)}{\pi_j(1-\pi_j)} \leqslant e^{\gamma} \tag{6}$$

The two individuals have the same probability of participating in a training program if $e^{\gamma} = 1$, which implies no unobserved selection bias. If for instance $e^{\gamma} = 2$, then two individuals who appear similar on the x-vector differ in their relative odds of participating by a factor of two. If e^{γ} close to 1 changes the inference about the training effect, then estimated training effects are said to be sensitive to unobserved selection bias. However, if a large value of e^{γ} does not alter inferences about the training effect, the study is not sensitive to selection bias.

For fixed $e^{\gamma} \geqslant 1$ and $u \in \{0, 1\}$ it can be shown under the null hypothesis that the test-statistics Q_{M-H} can be bounded by two known distributions, see Rosenbaum (1995, Prop. 4.3, ğ4.4.1). We will give these bounds for Q_{M-H} for each stratum, and for the test-statisitics of no overall effect of training.

For $e^\gamma = 1$ the upper and lower bounds are both equal to the value of the estimated test-statistics reported in Table 3. for increasing e^γ the bounds move apart reflecting uncertainty about the test-statistics in the presence of unobserved selection bias.

Let Q^+_{M-H} be the test-statistics given that we have overestimated the training effect in Table 3, and Q^-_{M-H} the test-statistics given that we have underestimated the training effect for different values of unobserved selection bias. Given no unobserved selection bias the expected number of employed participants under the null hypothesis of no training effect is given by $n_{1s} \cdot (y_s/n_s)$, where n_{1s} is the number of training participants, y_s the number of employed individuals, and n_s the number of individuals in stratum s. Given selection bias, the probability of participating is not the same for everyone in stratum s, but rather depends on unobserved factors u. The two bounds for the overall effect of training are given by

$$Q^+_{M-H} = \frac{\left[\sum_{s=1}^{S}(y_{1s} - \tilde{E}^+_s)\right]^2}{\sum_{s=1}^{S} Var(\tilde{E}^+_s)} \tag{7}$$

and

$$Q^-_{M-H} = \frac{\left[\sum_{s=1}^{S}(y_{1s} - \tilde{E}^-_s)\right]^2}{\sum_{s=1}^{S} Var(\tilde{E}^-_s)} \tag{8}$$

where $\tilde{E}_s$ and $Var(\tilde{E}_s)$ are the large sample approximations to the expectation and variance of the number of trainees employed, y_{1s}, when the N-vector u has binary coordinates, $u_{si} = 1$ or $u_{si} = 0$ for each (s, i), and for given γ. The calculation of $\tilde{E}_s$ and $Var(\tilde{E}_s)$ is based on the extended hypergeometric distribution. For details about this distribution see for instance (Johnson, Kotz and Kemp 1992) ğ6.11. The large sample approximation of $\tilde{E}^+_s$ is the unique root of the following quadratic equation

$$\tilde{E}^2_s(e^\gamma - 1) - \tilde{E}_s[(e^\gamma - 1) + (n_{1s} + y_s) + n_s] + e^\gamma y_s n_{1s}) \tag{9}$$

with the addition of $max(0, y_s + n_{1s} - n_s) \leqslant \tilde{E}_s \leqslant min(y_s, n_{1s})$ to decide which root to use. $\tilde{E}^-_s$ is determined by replacing e^γ by $\frac{1}{e^\gamma}$. The large sample approximation of the variance is given by

$$Var(\tilde{E}_s) = \left(\frac{1}{\tilde{E}_s} + \frac{1}{(y_s - \tilde{E}_s)} + \frac{1}{n_{1s} - \tilde{E}_s)} + \frac{1}{n_s - y_s - n_{1s} - \tilde{E}_s)}\right)^{-1} \qquad (10)$$

The upper and lower bounds for each stratum are found by removing $\sum_{s=1}^{S}$ in equations (7) and (8). Equation (9) and (10) are due to (Stevens 1951).

Results

Table 4 shows the sensitivity of the test-statistics for $e^{\gamma} = 1.25$, $e^{\gamma} = 1, 5$ and $e^{\gamma} = 2$ together with the test-statistics for $e^{\gamma} = 1$, which in the model implies that there is no unobserved selection bias. For $e^{\gamma} = 1.25$, even though participants and nonparticipants are equally distributed in terms of observed background variables, they differ in terms of unobserved variables in the data. If we compare two individuals with the same x-vector, $e^{\gamma} = 1.25$ implies that they differ in their odds of participating in a training program by a factor of 1.25, or 25 percent. For $e^{\gamma} = 2$ then two individuals that have the same x-vector differ in their odds of participating in a training programme by a factor of 2, or 100 percent, which must be considered to be a very large number given that we have adjusted for many important observed back-

TABLE 4
Sensitivity Analysis for e = 1.25, e = 1.5 and e = 2

Group	*Chi-squared for e = 1*	*Bounds for Chi-squared, e = 1.25*	*Bounds for Chi-squared, e = 1.5*	*Bounds for Chi squared, e = 2*
1	17.62**	10.76 – 26.50**	6.45 – 35.28**	1.92 – 52.34
2	8.70**	3.95 – 15.47**	1.45 – 22.52	0.00 – 36.70
3	14.09**	7.85 – 22.40**	4.15 – 30.71**	0.68 – 46.96
4	0.14	(0.38) – 1.88	(2.05) – 4.78	(7.43) – 12.17
5	15.61**	9.28 – 23.88**	5.36 – 32.04**	1.38 – 47.90
6	3.09	0.74 – 7.12	0.02 – 11.69	(1.04) – 21.00
7	0.45	(2.70) – 0.08	(5.96) – 1.16	(13.84) – 5.44
8	0.89	0.00 – 3.56	(0.57) – 7.08	(3.86) – 15.23
9	0.70	(0.01) – 3.09	(0.69) – 6.32	(4.08) – 13.91
10	0.45	(2.67) – (0.57)	(5.91) – 1.17	(13.75) – 5.45
11	0.59	(2.69) – 0.03	(6.18) – 0.88	(13.94) – 4.65
12	1.69	(4.91) – (0.16)	(8.84) – 0.11	(17.57) – 2.28
Total	20.35**	1.56 – 62.79	(2.76) – 117.6	(12.21) – 219.8

Note. **Indicates that the training effect is not sensitive to selection bias. Numbers in parenthesis means that the training effect is nonpositive. For a one-sided test of training effect, where the alternative is a positive training effect, these numbers should be ignored or set to zero. In no cases are the training effect under the alternative hypothesis of a negative training effect insensitive to selection bias.

ground characteristics. The theory of mixture models suggests that γ itself can be bounded, see (Lindsay 1995). However, incorporating the mixture structure into the framework of matching is not a trivial task but suggests a fruitful future research agenda.

The bounds given in Table 4 for different e^{γ} can be interpreted in the following way: If we have a positive (unobserved) selection, in the sense that if those most likely to participate, given that they have the same *x*-vector, also have higher employment rates, then the estimated training effects in Table 3 overestimate the true training effects. The reported chi-square statistics is then too high, and should be adjusted downwards, thus $Q^{+}_{M-H} \leqslant Q_{M-H}$. If we have a negative (unobserved) selection, in the sense that if those most likely to participate, given that they have the same *x*-vector, have the lowest employment rates, then the estimated training effects in Table 3 underestimate the true training effect, and the test-statistics should be adjusted upwards. This should not be confused with the observed selection process discussed earlier. In Section III we found that there is positive (observed) selection into training programmes. However, unobserved selection need not follow the same pattern as observed selection. In our sensitivity analyses we take unobserved selection to the extremes. If the estimates are sensitive to selection bias, then the training effect may still be positive, but it may also be zero, or even negative; it depends on the direction and magnitude of the selection bias. A sensitivity analysis shows how biases might alter inferences. However, it does not indicate whether biases are present or what magnitudes are plausible.

Table 4 shows that most of the estimated training effects are sensitive to unobserved selection bias. However, the positive training effects estimated for groups 1, 3, 5, and to some degreee group 2, are partly robust to selection bias (the numbers indicated by **). The results for groups 1, 3 and 5 indicate a positive training effect for these groups even if we allow trainees and comparisons to differ with as much as 50 percent in terms of unobserved characteristics. The positive overall effect of 6.3 percentage points is however sensitve to selection bias. If persons with a high value of *u* are over represented in training programs, then the estimated training effect of 6.3 percentage points overestimates the true training effect, and the true effect is not different from zero. If those who have a low value of *u* are over represented, the estimated training effect underestimates the true training effect, and the true effect is highly significant.

V Concluding Remarks

In this paper we discuss selection bias in training models, and how sensitivity analysis can be used to evaluate the intrinsic uncertainty of estimated training effects. We illustrate the method by comparing empoloyment outcomes of trainees and nonparticipants with the same probabilities of training. Even though the estimated overall training effect is positive and significantly different from zero, given that we have adjusted for observed selection bias, it is sensitive to selection bias, and thus must be interpreted with caution. However, the positive effect for individuals with the lowest propensity scores is not sentitive to a 50 percent difference in the probability of training for comparable individuals. This means that even if there is 'cream-skimming' in the vocational rehabilitation sector, both through observed and unobserved variables, in the sense that persons participating in training programmes are those most likely to be employed even without training, there is a positive effect of training for individuals with a low propensity score. Given constant costs across training programmes, and that unobserved 'creaming' does not change the rank ordering of the propensity score, there is scope for improving the effectiveness and redistribution effect of the VR sector by reversing the selection rule from 'cream-skimming' to 'bottom fishing', although this strategy potentially may increase the stigma of VR training programmes, and thus in the long run have a decremental effect.

Date of Receipt of Final Manuscript: March 2000

References

Aakvik, Arild and Risa, Alf Erling (1998). 'Success through Selection in Norwegian Rehabilitation Programs?,' *Working Paper*, *University of Bergen*.

Ahn, Hyungtaik and Powell, James L. (1993). 'Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism,' *Journal of Econometrics*, Vol. 58(1–2), pp. 3–29.

Ashenfelter, Orley (1978). 'Estimating the Effects of Training Programs on Earnings,' *Review of Economics and Statistics*, Vol. 60, pp. 47–57.

Ashenfelter, Orley and Card, David (1985). 'Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs,' *Review of Economics and Statistics*, Vol. 67, pp. 648–60.

Bassi, Laurie J. (1983). 'The Effect of CETA on the Postprogram Earnings of Participants,' *Journal of Human Resources*, Vol. 18, pp. 539–56.

Bell, Stephen H., Orr, Larry L., Blonquist, John D. and Cain, Glen G. (1995). *Program Applicants as a Comparison Group in Evaluating Training Programs*, W. E. Upjohn Institute for Employment Research, Kalamazoo, Michigan.

Card, David and Sullivan, Daniel (1988). 'Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment,' *Econometrica*, Vol. 56, pp. 497–530.

Cochran, W. G. (1968). 'The Planning of Observational Studies of Human Populations,' *Journal of the Royal Statistical Society*, *Ser. A*, Vol. 128, pp. 234–55.

Ford, Margaret (1993). 'Attføring til Arbeid? (Rehabilitation to Work?),' *INAS-rapport*, Vol. 93.

Ham, John C. and LaLonde, Robert J. (1996). 'The Effects of Sample Selection and Initial

Conditions in Duration Models: Evidence from Experimental Data,' *Econometrica*, Vol. 64, pp. 175–205.

Heckman, James J. (1979). 'Sampe Selection Bias as a Specification Error,' *Econometrica*, Vol. 47, pp. 153–61.

Heckman, James J. and Honoré, Bo (1990). 'Empirical Content of the Roy Model,' *Econometrica*, Vol. 58, pp. 1121–49.

Heckman, James J. and Robb, Richard (1985). 'Alternative Methods for Evaluating the Impact of Interventions: An Overview,' *Journal of Econometrics*, Vol. 30, pp. 239–67.

Heckman, James J. and Hotz, Joseph V. (1989). 'Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training,' *Journal of the American Statistical Association*, Vol. 84, pp. 862–74.

Heckman, James J., Ichimura, Hidehiko, Smith, Jefferey and Todd, Petra (1998). 'Characterizing Selection Bias Using Experimental Data,' *Econometrica*, Vol. 66, pp. 1017–98.

Johnson, Norman Lloyd, Kotz, Samuel and Kemp, Adienne W. (1992). *Univariate Discrete Distributions*, Wiley, New York.

Lindsay, Bruce G. (1995). *Mixture Models: Theory, Geometry, and Applications*, Hayward, Calif.: Institute of Mathematical Statistics; Alexandria, Va.: American Statistical Association.

Mantel, N. and Haenszel, W. (1959). 'Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease,' *Journal of the National Cancer Institute*, Vol. 22, pp. 719–48.

Newey, Whitney K., Powell, James L. and Walker, James R. (1990). 'Semiparametric Estimation of Selection Models: Some Empirical Results,' *American Economic Review Papers and Proceedings*, Vol. 80, pp. 324–28.

Quandt, Richard (1988). *The Econometrics of Disequilibrium*, Oxford: Blackwell.

Rosenbaum, Paul R. (1995). *Observational Studies*, Springer-Verlag, New York.

Rosenbaum, Paul R. and Rubin, Donald B. (1984). 'Reducing Bias in Obersvational Studies Using Subclassification on the Propensity Score,' *Journal of American Statistical Association*, Vol. 79, pp. 516–24.

Roy, Andrew D. (1951). 'Some Thoughts on the Distribution of Earnings,' *Oxford Economic Papers*, Vol. 3, pp. 135–46.

RTV (1985). *Rikstrygdeverkets rundskriv*, RTV kom. 05-02 10/85.

Rubin, Donald B. (1986). 'Which Ifs Have Causal Answers?,' *Journal of the American Statistical Association*, Vol. 81, pp. 961–62.

Stevens, W. L. (1951). 'Mean and Variance of an Entry in a Contingency Table?,' *Biometrica*, Vol. 38, pp. 468–70.

Appendix

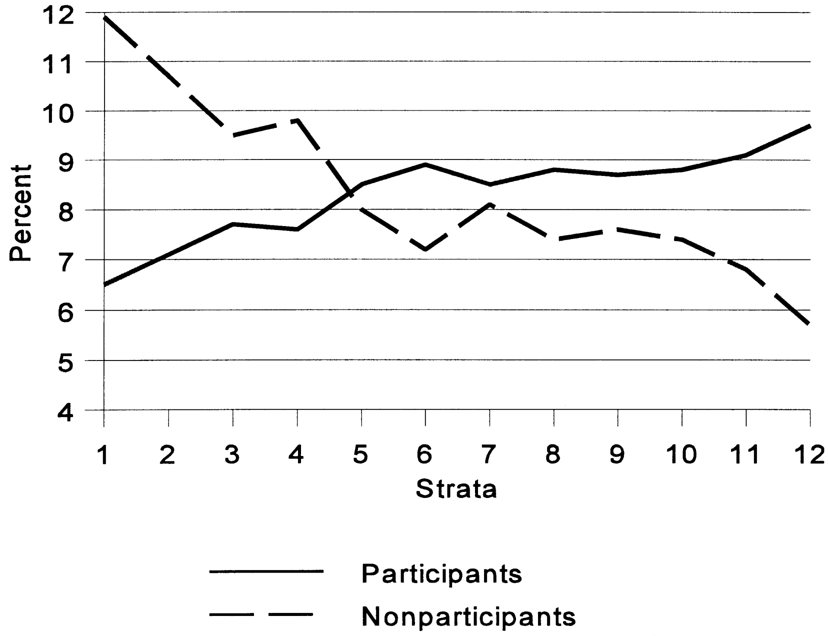

Figure 1. The percentage of participants and nonparticipants in different strata

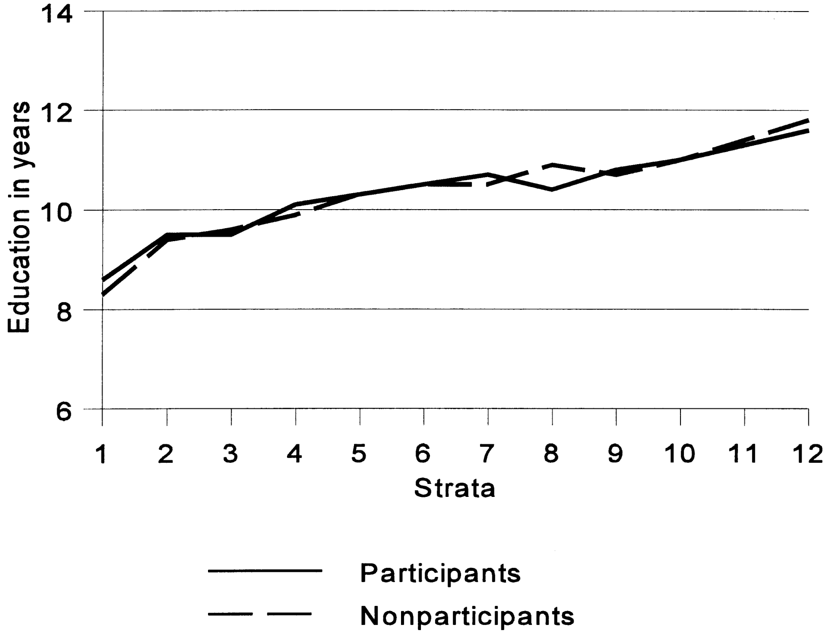

Figure 2. Mean education of participants and nonparticipants in different strata after matching

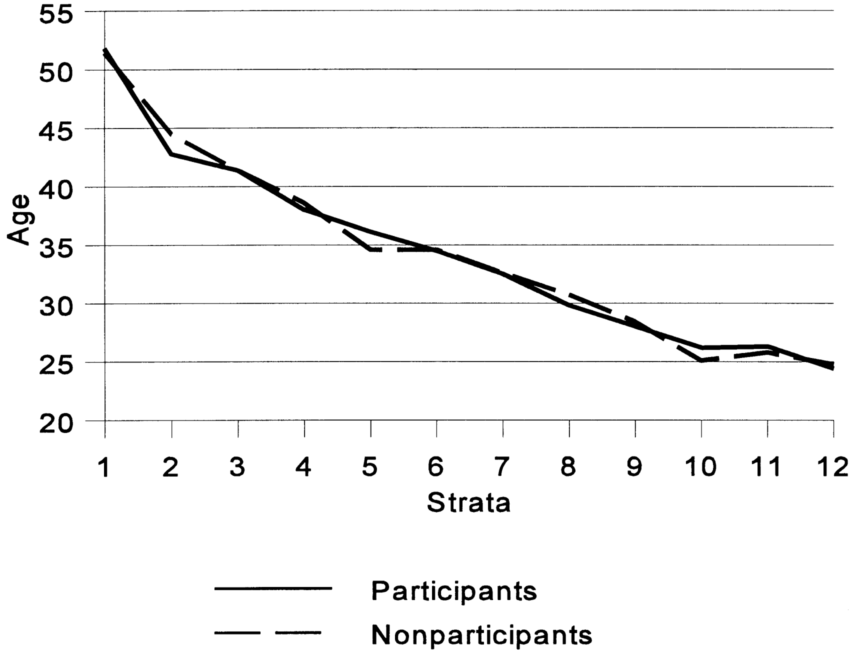

Figure 3. Mean age of participants and nonparticipants in different strata after matching

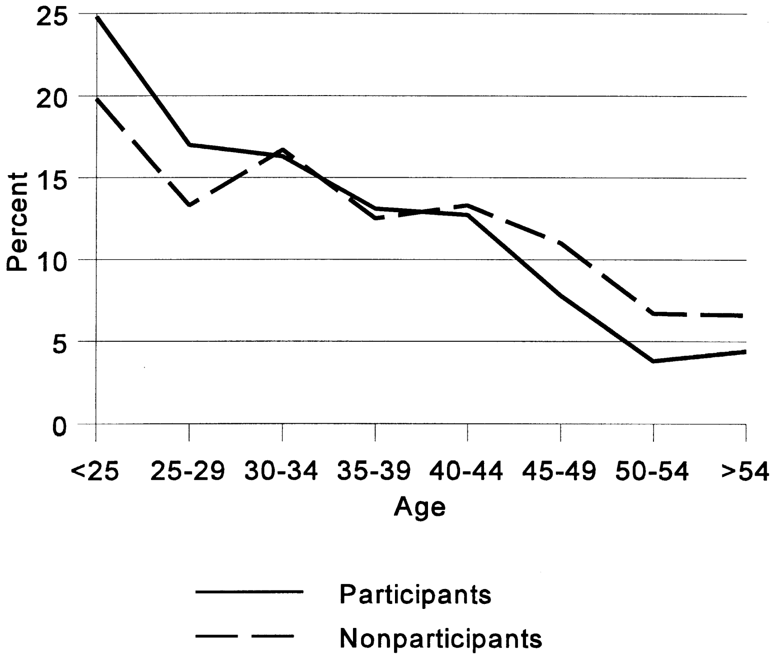

Figure 4. The distribution of age for the total sample

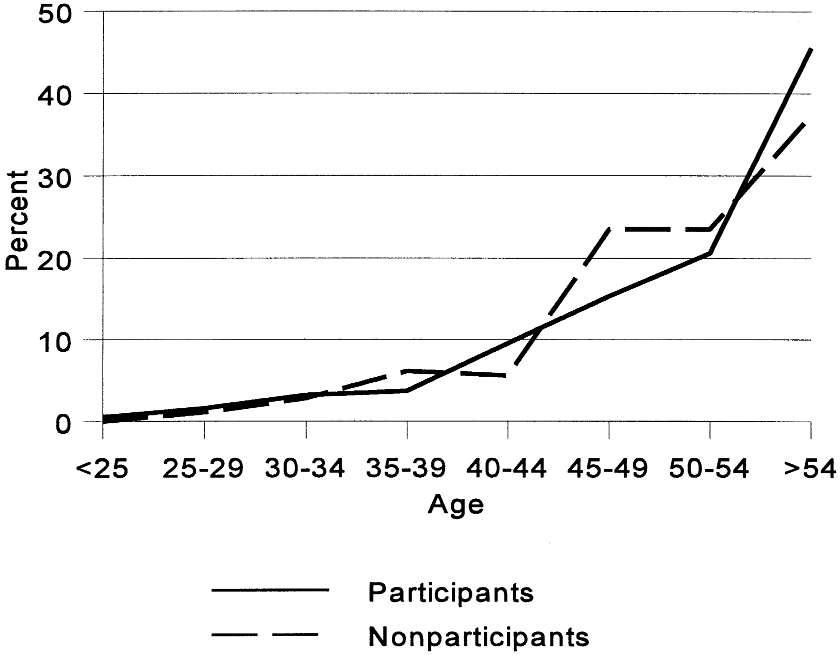

Figure 5. The distribution of age for stratum 1

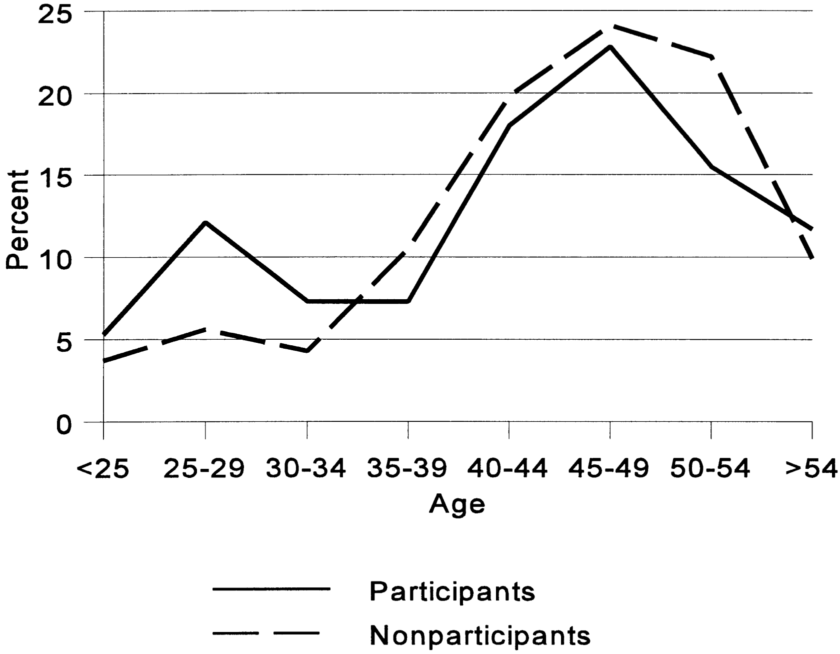

Figure 6. The distribution of age for stratum 2

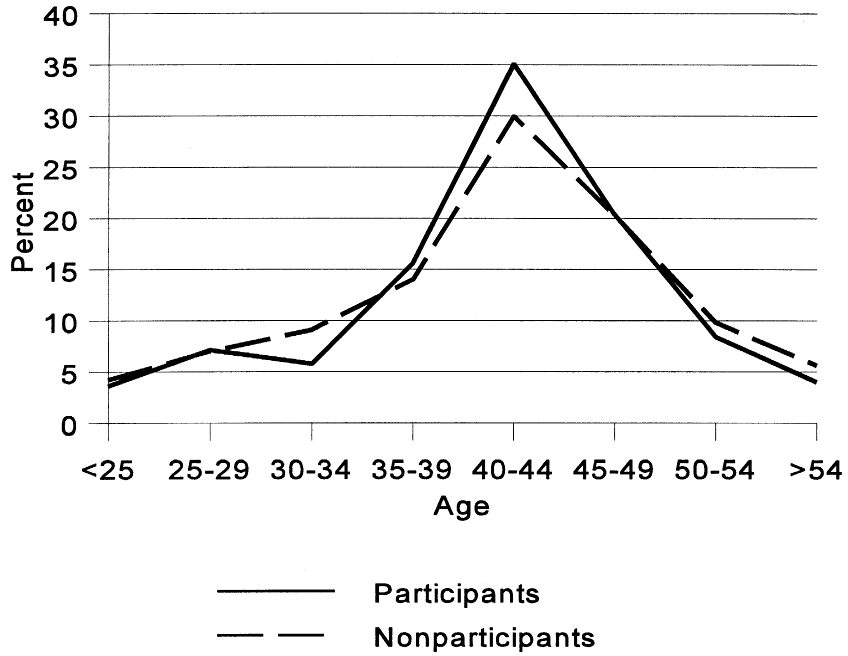

Figure 7. The distribution of age for stratum 3

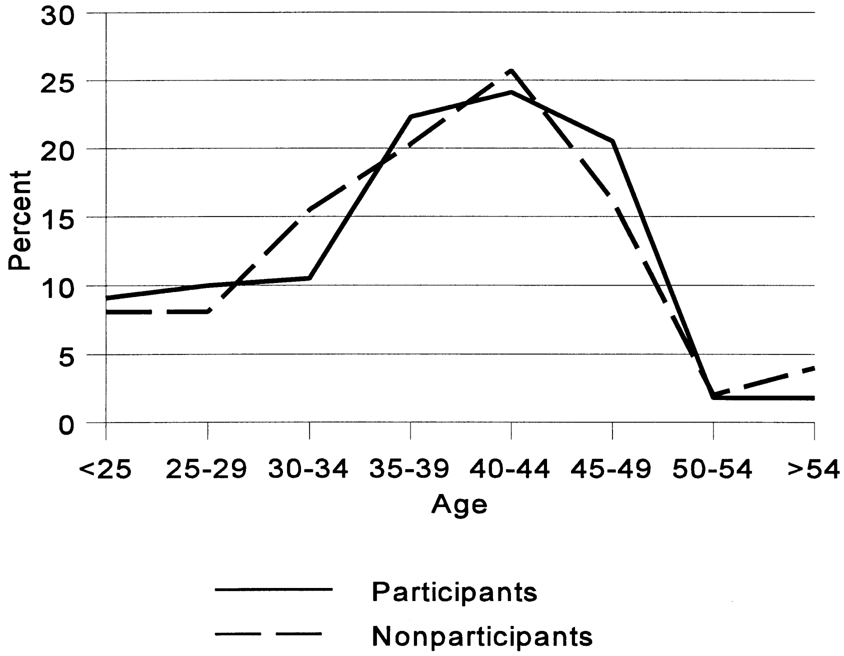

Figure 8. The distribution of age for stratum 4

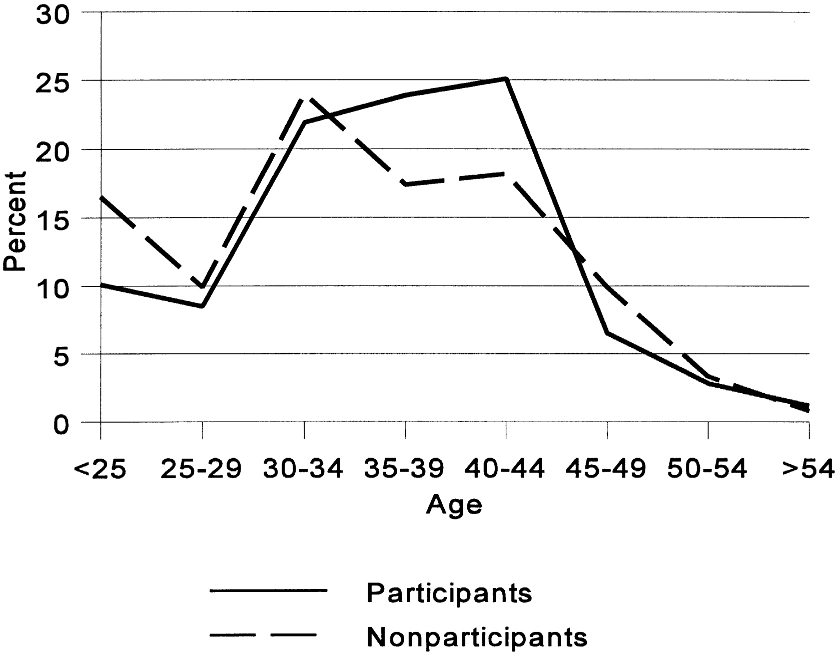

Figure 9. The distribution of age for stratum 5

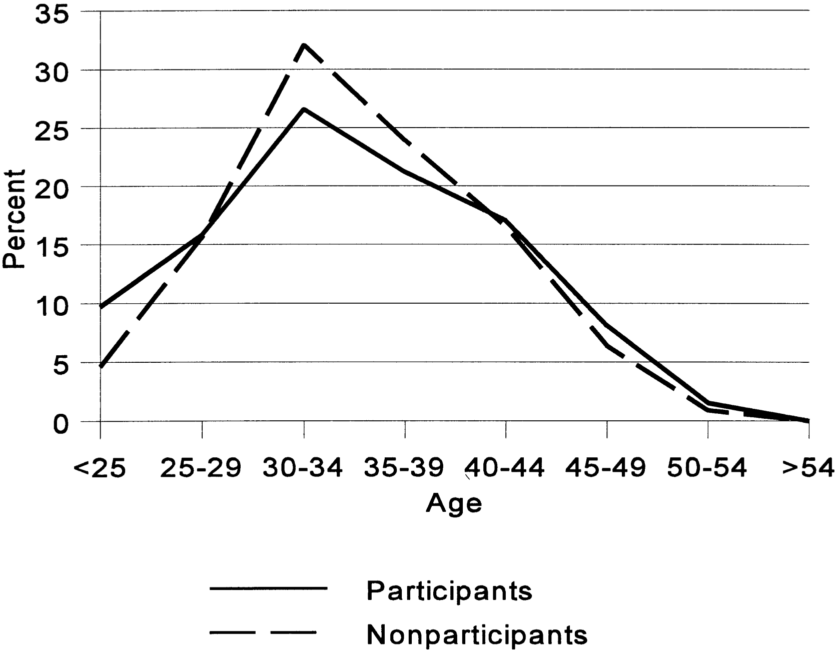

Figure 10. The distribution of age for stratum 6

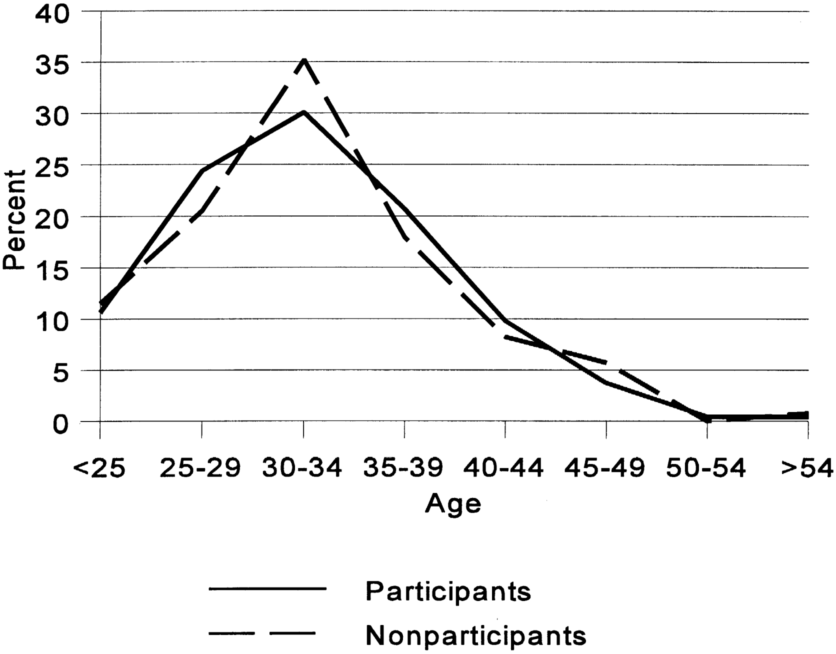

Figure 11. The distribution of age for stratum 7

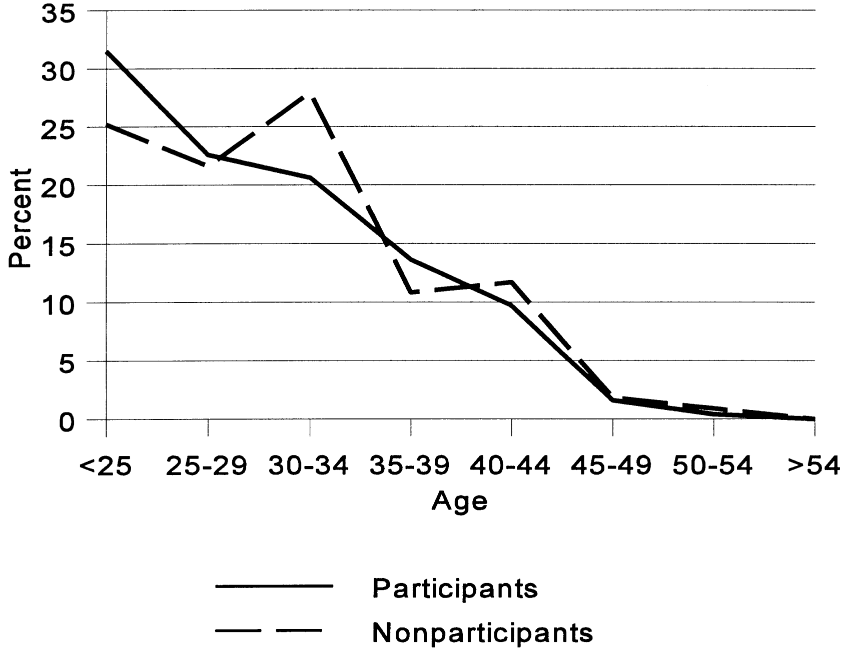

Figure 12. The distribution of age for stratum 8

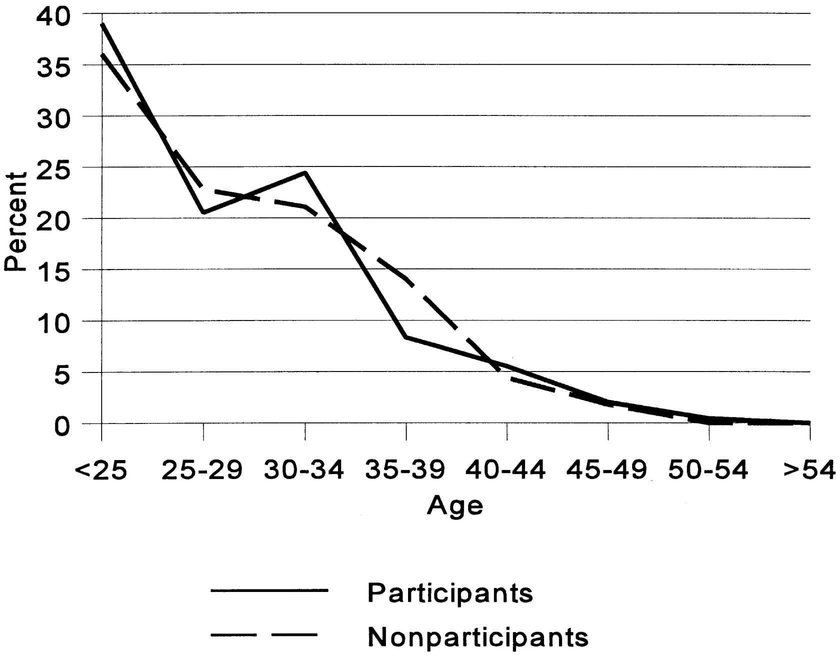

Figure 13. The distribution of age for stratum 9

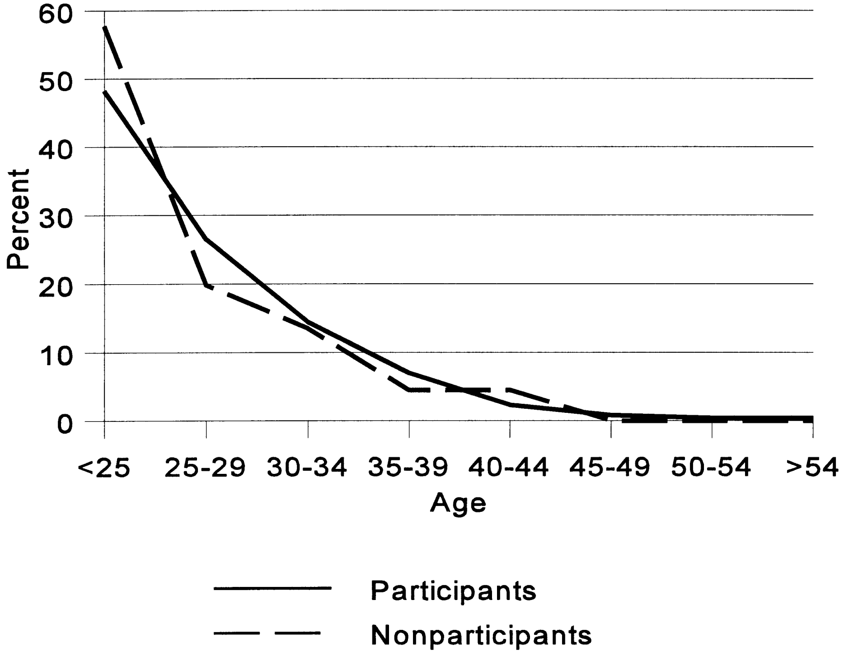

Figure 14. The distribution of age for stratum 10

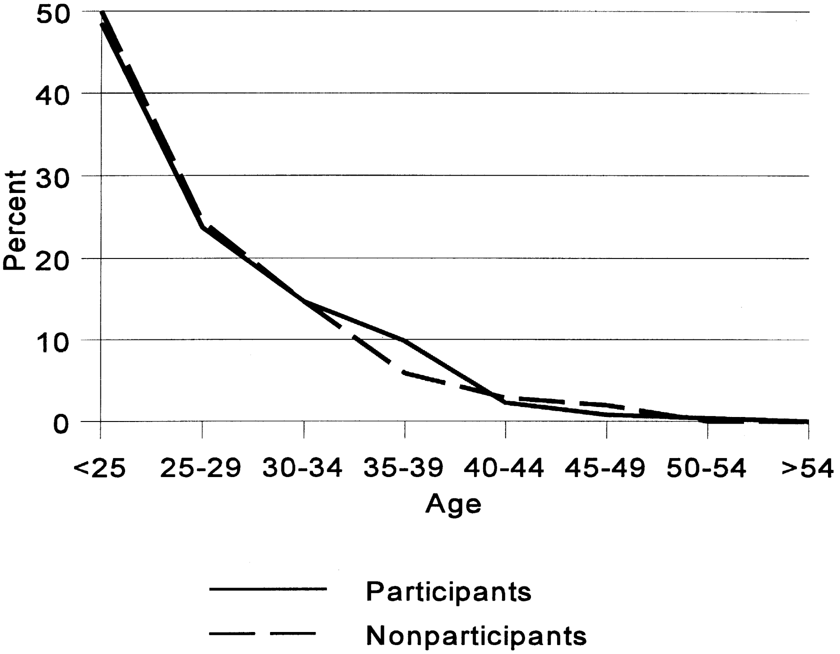

Figure 15. The distribution of age for stratum 11

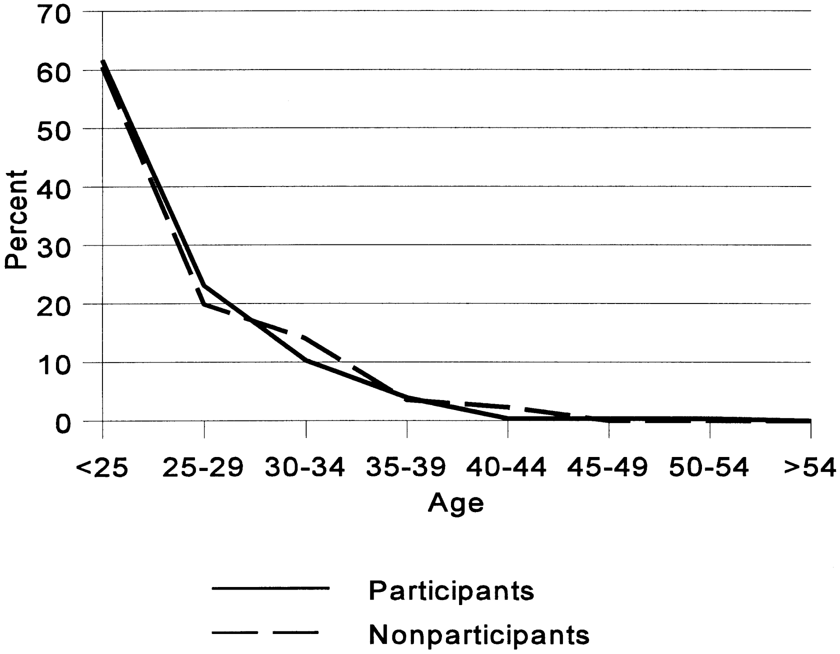

Figure 16. The distribution of age for stratum 12