Active Labor Market Policy Evaluations: A Meta-analysis

David Card
UC Berkeley

Jochen Kluve
RWI - Essen

Andrea Weber
UC Berkeley and RWI-Essen

February 2009

*We thank the many authors who responded to our survey for their co-operation and assistance. We also thank participants at seminars at the University of Chile and the Inter-America Development Bank for comments and suggestions. Our research was funded by the Center for Labor Economics at UC Berkeley and by RWI-Essen.

Active Labor Market Policy Evaluations:
A Meta-Analysis

ABSTRACT

This paper presents a meta-analysis of recent microeconometric evaluations of active labor market policies. Our sample consists of 199 program estimates drawn from 97 studies conducted between 1995 and 2007. In about one-half of these cases we have both a short-term impact estimate (for a one-year post-program horizon) and a medium-term estimate (two-year horizon). We characterize the program estimates according to the type and duration of the program, the characteristics of the participants, and the evaluation methodology. Heterogeneity in all three dimensions affects the likelihood that an impact estimate is significantly positive, significantly negative, or statistically insignificant. Comparing program types, subsidized public sector employment programs have the least favorable impact estimates. Job search assistance programs have relatively favorable short-run impacts, whereas classroom and on-the-job training programs tend to show better outcomes in the medium-run than the short-run. Programs for youths are less likely to yield positive impacts than untargeted programs, but there are no large or systematic differences by gender. Methodologically, we find that the outcome variable used to measure program effectiveness matters. Evaluations based on registered unemployment durations are more likely to show favorable short-term impacts. Controlling for the outcome measure, and the type of program and participants, we find that experimental and non-experimental studies have similar fractions of significant negative and significant positive impact estimates, suggesting that the research designs used in recent non-experimental evaluations are unbiased.

David Card
549 Evans Hall, #3880
UC Berkeley
Berkeley CA 94705
USA
and NBER

Jochen Kluve
RWI-Essen
Hohenzollernstr. 1-3
D-45128 Essen
GERMANY

Andrea Weber
549 Evans Hall, #3880
UC Berkeley
Berkeley CA 94705
USA
and RWI-Essen

The effectiveness of active labor market policies – including subsidized employment, training, and job search assistance – has been a matter of vigorous debate over the past half century.[1] While many aspects of the debate remain unsettled, some progress has been made on the key question of how participation in an active labor market program (ALMP) affects the labor market outcomes of the participants themselves.[2] Progress has been facilitated by rapid improvements in data and methodology, and by a growing institutional commitment to evaluation in many countries, and has resulted in an explosion of professionally authored microeconometric evaluations. In their influential review Heckman, Lalonde and Smith (1999) summarize approximately 75 microeconometric evaluation studies from the U.S. and other countries. A more recent review by Kluve (2007) includes nearly 100 separate studies.

In this paper we attempt to synthesize some of the main lessons in the recent microeconometric evaluation literature, using a new and comprehensive sample of program estimates from the most recent generation of studies. Our sample is derived from responses to a survey of 361 academic researchers affiliated with the Institute for the Study of Labor (IZA) and the National Bureau of Economic Research (NBER) in Spring 2007. These researchers and their colleagues authored or co-authored a total of 97 studies of active labor market policies between

---

[1]In the U.S., for example, the direct public sector employment programs initiated by the Works Progress Administration in 1935 were immediately controversial. The issue of whether the government should provide training to unemployed workers was also debated prior to World War II. See Wilson (2004), who argues that the GI Bill (the Servicemen's Readjustment Act of 1944) represented the first major step toward the institutionalization of active labor market programs in the U.S.

[2]A key unsettled question is whether ALMP's affect the outcomes of those who do not participate, via displacement or other general equilibrium effects. See Johnson (1976) for an early but informative general equilibrium analysis of public sector employment programs, and Calmfors (1994) for a more recent critique, focusing on the European experience of the 1980s and early 1990s.

1995 and 2007 that meet our inclusion criteria. We conduct a meta-analysis using a sample of 199 program estimates extracted from these studies.

Importantly, for about one-half of the sample we have both a short-term impact estimate – measuring the effect on participant outcomes approximately one year after the completion of the program – and a medium-term estimate giving the effect approximately 2 years after completion. We also have a longer-term impact (for about 3 years after completion) for about one-quarter of the programs. These estimates allow us to compare shorter- and longer-term effects of different types of programs, and assess the extent to which the program effects fade or grow over time.

We classify the estimates by whether the post-program impact on the participants is found to be significantly positive, statistically insignificant, or significantly negative. This classification, while admittedly crude, allows us to make comparisons across studies that use very different dependent variables – ranging from the duration of time in registered unemployment to average quarterly earnings – and that are obtained from very different institutional environments. Our main analysis uses an ordered probit framework, assuming that program effectiveness is a partially-observed latent random variable. We test this assumption by fitting separate probit models for the occurrence of significantly positive or significantly negative impact estimates and find reasonable support for the latent index assumption.

Our meta analysis model assumes that measured ALMP effectiveness depends on the type and duration of the program, the characteristics of the participants, and the evaluation methodology. Consistent with earlier summaries, we find that subsidized public sector employment programs are relatively ineffective, whereas job search assistance (JSA) programs have generally favorable impacts, especially in the short run. Classroom and on-the-job training

programs are not especially favorable in the short-run, but have more positive relative impacts after two years. Comparing across different participant groups, we find that programs for youths are less likely to yield positive impacts than untargeted programs, although in contrast to some earlier reviews we find no large or systematic differences by gender. We also find that evaluations based on the duration of time in registered unemployment are more likely to show favorable short-term impacts than those based on direct labor market outcomes (employment or earnings).

An important theme in the microeconometric evaluation literature is the difficulty of controlling for selection biases that may lead to specious positive or negative program effects.[3] This concern led observers in the 1980s to call for randomized evaluations of active labor market programs (e.g., Ashenfelter, 1987). Ultimately a significant number of controlled randomized trials have been conducted in the U.S. and elsewhere, and 18 of the estimates in our meta analysis sample (9%) are based on a randomized design. This feature allows us to conduct a comparison between the results of experimental and non-experimental evaluations, while controlling for the nature of the program and its participants. Controlling for the program type and composition of the participant group, we find that the differences between the experimental and non-experimental impact estimates are small and statistically insignificant ($t<0.5$), suggesting that the research designs used in recent non-experimental evaluations are not significantly biased relative to the "gold standard" of an experimental design.

The next section of the paper describes the procedures we used to collect a sample of

[3]See, e.g., Ashenfelter (1978), Ashenfelter and Card (1985), Heckman and Robb (1985), Lalonde (1986), Heckman, Ichimura, Smith and Todd (1998), and Heckman, Lalonde and Smith (1999).

recent microeconometric evaluation studies, and the criteria we used for including a study in our analysis sample. Section III presents a descriptive overview of the program estimates we extracted from the included studies. Section IV presents our main meta analysis results. Section V concludes the paper.

II. Assembling a New Sample of ALMP Program Estimates

a. Initial Survey of Researchers

To develop a comprehensive sample of recent ALMP evaluations we conducted a survey of academic researchers affiliated with two leading research networks: the Institute for the Study of Labor (IZA) and the National Bureau of Economic Research (NBER).[4] We obtained the email list for IZA research fellows who had indicated an interest in the program area "Evaluation of labor market programs", and the list for associates of the NBER Labor Studies program. We sent each network member a personally addressed email with a cover letter explaining that we were trying to collect all the recent (post-1990) microeconometric program evaluation studies that they or their students or colleagues had written. In addition, we attached a questionnaire that we asked them to complete for each study they had produced.[5]

Our list of IZA fellows was extracted on January 25, 2007, and contained a total of 232 names (excluding the three of us). We emailed the survey on February 21st, 2007. Three email

[4]The formal meta analysis literature stresses the importance of collecting a comprehensive sample of studies (e.g., Higgins and Green, 2008). Much of that literature is concerned with the problem of collecting unpublished studies or studies published in non-journal outlets (so-called "grey literature"). We believe that by surveying the producers of relevant studies we have largely avoided this problem. In fact, only 45% of the program estimates in our sample are derived from published studies.

[5]The questionnaire is available on request.

addresses turned out to be invalid, but we were able to identify a correct address for two, yielding a final IZA-based sample of 231. We followed a similar procedure for affiliates of the NBER Labor Studies Program, extracting names and email addresses on March 20, 2007. After eliminating names of those already on the IZA list, we emailed our survey to 130 NBER associates on March 22, 2007. In our email we asked respondents to identify colleagues and students working on microeconometric ALMP evaluations. We were forwarded a total of 14 additional names (and emails) who constituted a third part of our sample frame.

Table 1 summarizes the responses to our survey. The overall response rate across the 375 researchers we ultimately contacted was 53%. The response rate was somewhat higher for IZA fellows than NBER Associates, and was quite high among the small group of 14 additional researchers referred to us by the original sample members.[6] Among respondents, 57% reported that they had no relevant studies to contribute. The remaining group of 84 researchers returned a total of 156 separate studies that form the basis for our sample.

b. Selection of Studies

The next step in our process was to define the types of active labor market programs and the types of evaluation methods that we would consider "in scope" for our meta-analysis. We imposed four restrictions on the kinds of programs to be included. First, the ALMP had to be one of the following types:

-classroom or on-the-job training

[6]The higher response rate for IZA members may be due to the fact that the IZA list is made up researchers who self-identified as interested in labor market program evaluations.

-job search assistance or sanctions for failing to search[7]

-subsidized private sector employment

-subsidized public sector employment

or a combination of these types. Second, we restricted the definition of private or public employment subsidies to include only individual-level subsidies. That is, we excluded firm-level subsidy programs that allow employers to select the individuals whose jobs are subsidized. Third, we restricted attention to time-limited programs. This criterion eliminates open-ended entitlements like general education subsidies and child care programs. Finally, we decided to focus on programs with an explicit "active" component. Thus, we excluded purely financial programs, such as manipulations of the benefits available to participants in unemployment insurance, welfare or disability programs.

In terms of methodology we decided to limit our attention to well-documented empirical evaluation studies based on individual microdata. Thus, we exclude purely theoretical studies and survey articles, as well as studies that use regional or national time series data. Finally, we consider only those evaluations that have an explicit comparison or control group of individuals who were not subject to the program (or who entered the program at a later date).

Applying these rules, we eliminated 33 of the originally submitted studies that did not meet our ALMP program requirements and 18 that did not meet our methodological criteria. We also eliminated 8 studies that were written in a language other than English, or had substantial overlap with other studies included in the sample (e.g., earlier versions of the same study), or were otherwise incomplete. The remaining 97 studies (=156−33−18−8) form the basis for our

[7]A couple of programs are actually based on the threat of assignment to a program, which we interpret as a form of sanctions. See e.g., Hagglund (2007).

empirical analysis.

c. Extraction of Program Estimates and Other Information

Having identified a set of studies, our next step was to extract information about the program and participants analyzed in each study, and the estimated program impact(s). Although we initially intended to gather this information from the questionnaires distributed in our email survey, we were unable to use these forms because only 38% of authors attempted to complete the questionnaire. Ultimately, we decided to extract the information ourselves.[8]

Many variables were relatively straightforward to collect, including the type of program, the age and gender of the participant population, the type of dependent variable used to measure the impact of the program, and the econometric methodology. It proved more difficult to find information on the comparability of the treatment and control groups, and to gauge the plausibility of the econometric methodology. Despite the emphasis that prominent methodologists have placed on documenting the degree of "overlap" between the characteristics of the participants and the comparison group, relatively few studies present detailed information on the pre-program characteristics of the participants and the comparison group.[9] Another (surprising) fact is that very few studies provide information on program costs. We decided to use average program duration as a rough proxy for the size of the investment represented by the

[8]We found that even graduate level research assistants had some difficulty understanding the studies, so we each read and classified about one-third of the studies. We acknowledge that there are likely to be some (potentially large) measurement errors and errors of interpretation in the extraction of information from the studies.

[9]See e.g., Heckman, Ichimura, Smith and Todd (1998), and Heckman, Ichimura and Todd (1998).

program.

The most difficult task, however, proved to be the development of a standardized measure of program impact that could be compared across studies. This is mainly due to the wide variation in methodological approaches in the recent ALMP literature. For example, about one-third of the studies in our sample report treatment effects on the hazard rate from registered unemployment. Very rarely do these studies include the information needed to infer the implied impacts on the probability of employment at some date after the completion of the program, which is the most commonly used outcome variable (45% of studies).

Faced with such a diverse set of outcome measures we abandoned the preferred meta analysis approach of extracting a standardized "effect size" estimate from each study.[10] Instead, we classified the estimates into three qualitative categories: significantly positive, insignificantly different from zero, and significantly negative.[11] We also classified the estimates based on a rough assessment of the elapsed time since completion of the program into *short-term impacts*, measured in the first 12 months after program participation, *medium-term impacts*, measured 12 to 24 months after program participation, or *long-term impacts*, measured more than two years after program participation.

---

[10]See Hedges and Olkin (1985). The U.S. Department of Education's *What Works Clearinghouse*, for example, presents meta analytic summaries of the effectiveness of education interventions using estimated impacts on student test scores, normalized by the standard deviation of the test used in the study.

[11]This is slightly different than the so-called "vote count" approach of classifying estimates by whether they are significantly positive or not because estimates in our context can be significantly negative. We discuss some of the potential concerns with our methodology in Section III, below. Vote counting is especially problematic when individual studies have low power (so an insignificant outcome is likely, even when the true effect is non-zero) and/or when statistically insignificant results are less likely to be written up and published. See Card and Krueger (1995) for an example of the latter problem in the literature on minimum wages.

Many studies in our sample report separate impacts for different programs types (e.g., job training versus private sector employment) and/or for different participant subgroups. Whenever possible, we extracted separate estimates for each program type and participant subgroup combination, classifying participant groups by gender (male, female, or mixed) and age (under 25, 25 and older, or mixed). Overall, we extracted a total of 199 "program estimates" (estimates for a specific program and participant group) from the 97 studies in our sample.[12] For many of the program/subgroup combinations we have both a short-term impact and a medium- and/or long-term impact. Specifically, for 108 program/subgroup combinations we have a short-term and medium term program impact. For 48 program/subgroup combinations we have a short-term and a long-term impact estimate.

d. Sample Overview

Table 2 shows the distribution of our sample of program estimates by the latest "publication date" of the study (panel a) and by country (panel b).[13] The studies included in our sample are all relatively recent: 90% come from articles or working papers published in 2000 or later, and 45% from papers published in the last 3 years. The program estimates cover a wide range of countries (26 in total), with the largest numbers of estimates for evaluations from Germany (45 estimates), Denmark (26 estimates), Sweden (19 estimates) and France (14 estimates). Interestingly, only 10 estimates are from U.S. studies.

---

[12]56 studies contribute a single program estimate, 17 studies contribute 2 program estimates, and 24 studies contribute 3 or more program estimates.

[13]Note that 55% of the estimates are from unpublished studies. By "publication date" we mean the date on the study, whether published or not.

III. Descriptive Analysis

a. Program Types, Participant Characteristics, and Evaluation Methodology

Table 3 presents a summary of the characteristics of the types of programs and types of program participants represented in our sample of 199 program estimates. To aid in a discussion of the sample we find it useful to define three broad country groups that together represent about 70% of the program estimates. Countries in each group share many important institutional features and also tend to have similar design features in their active labor market programs. The largest group of estimates are from the German speaking countries (Austria, Germany, and Switzerland) with 67 program estimates (column 2 of Table 3). The second largest group are from the Scandinavian countries (Denmark, Finland, Norway, and Sweden) with 53 program estimates (column 3). A third distinct group are the "Anglo" countries (Australia, Canada, New Zealand, U.K. and U.S.). For this group - summarized in column 4 of Table 3 – we have 20 program estimates.

Rows 2a-2c of Table 3 illustrate a first important contrast between the three main country groups by showing the variation in the sources of ALMP participants. Overall, nearly 70% of the program estimates are for programs targeted at people who enter from registered unemployment. This is particularly likely for programs in the German-speaking countries, where 94% of the estimates are for participants from registered unemployment, but it is also generally true in the Scandinavian countries, and in most of Continental Europe. In contrast, in the Anglo countries only a small fraction (15%) of the estimates are for programs with participants drawn from the unemployment insurance program. In these countries, many training and subsidized employment

programs are targeted at long-term disadvantaged individuals who are enrolled via community outreach programs or from the welfare system.

Rows 3a-3f show the types of active labor market programs in our sample. Classroom and work experience training programs are the most common, particularly in the German-speaking countries, where 63% of the program estimates are for classroom or on-the-job training programs. Job search assistance programs are relatively uncommon in the German-speaking and Scandinavian countries but are relatively common in the Anglo countries.[14] Subsidized pubic and private employment programs together account for about 30% of our sample of program estimates, and are relatively evenly distributed across the three main country groups. Finally, combination programs are particularly common in Scandinavia, where people who remain in registered unemployment often are automatically assigned to some form of "active" program (see, e.g., Sianesi, 2004).

Rows 4a-4d show the distribution of program durations. In general, most active labor market programs are short, with a typical duration of 4-6 months. Programs tend to be somewhat longer in the German-speaking countries and shorter in the Anglo countries. The short duration of the programs suggests that at best they might be expected to have relatively modest effects on the participants – comparable, perhaps to the impact of an additional year of formal schooling. Our impression is that an impact on the order of a 5-10% permanent increase in labor market earnings (or a somewhat larger short-term impact) would be large enough to justify many of the

[14]In most countries people receiving unemployment benefits are eligible for some form of job search assistance, which we would not consider in scope for our review. The job search assistance programs included in our sample are special programs outside of these usual services (or in some cases provided to people who are not in registered unemployment).

programs on a cost-benefit basis.[15]

Rows 5 and 6 of Table 3 present data on the gender and age composition of the participant groups associated with the program estimates. Our reading is that very few of the programs themselves are targeted by gender: rather, in cases where gender-specific estimates are available it is because the authors have estimated separate impacts for the same programs on men and women. The situation with respect to age is somewhat different. Sometimes the programs are specifically targeted to younger workers (i.e., those under 21 or 25), whereas sometimes programs are available to all age groups but the analysts have limited their study to participants over the age of 24, or stratified by age.[16] In any case, the majority of the program estimates in our sample are for all ages and both genders.

Table 4 describes the features of the evaluation methods used in our sample. Apart from the randomized designs, there are essentially two main methodological approaches in the recent literature. One, which is widely adopted in the German-speaking and Anglo countries, uses longitudinal administrative data on employment and/or earnings for the participants and a comparison group (who are assigned to an simulated starting date for a potential program). Typically, the data set includes several years of pre-program labor market history, and propensity-score matching is used to narrow the comparison group to a sample whose observed characteristics and pre-program outcomes closely match those of the participants (see e.g., Gerfin

[15]Jespersen, Munch, and Skipper (2007) present a detailed cost-benefit analysis for various Danish programs, and conclude that subsidized public and private sector employment programs have a positive net social benefit, whereas classroom training programs do not.

[16]Sometimes the age restriction is imposed because the evaluation method requires 3-5 years of pre-program data, which is only available for older workers. Austria Germany and Switzerland have programs for younger workers that are incorporated into their general apprenticeship systems and are not typically identified as "active labor market programs."

and Lechner, 2002; Biewen, Fitzenberger, Osikominu, and Waller, 2007; Jespersen, Munch, and Skipper, 2007). In this type of study, the program effect is usually measured in terms of the probability of employment at some date after the completion of the program, although earnings can also be used. Over two thirds of the evaluations from the German-speaking and Anglo countries fit this mold, as do a minority (about 30%) of the evaluations from the Scandinavian countries.

The main alternative approach, which is widely used in the Scandinavian countries, is a duration model of the time to exit from registered unemployment – see e.g. Sianesi (2004). An important advantage of this approach is that it can be implemented using only data from the unemployment benefit system (i.e., without having access to employment records). The program effect is measured in terms of the difference in the duration of time to exit between the participants who entered a program at a certain date and those with similar characteristics and a similar history of unemployment who do not. In some studies the outcome variable is defined as the duration of time to exit *to a new job* while in others the exit event includes all causes.[17] Even in the former case, however, the program effect for entry at date $\boldsymbol{t}$ cannot be translated into an effect on future employment without specifying the risks of future program entry and the full set of future program effects.[18] Nevertheless, the sign of the treatment effect is interpretable, since a

[17]The implicit presumption is that people who exit for other reasons are not employed, so exit to employment is a "good" outcome but exit for other reasons is a "bad" outcome. As documented by Bring and Carling (2000), however, in the Swedish case nearly one-half of those who exit for other reasons are later found to be working.

[18]Richardson and Van den Berg (2002) consider a special case in which there is a constant hazard rate of program entry, and entry into a program at some time exerts a proportional effect on the hazard of exit to employment from then onward. This model can be used to predict impacts on future employment outcomes relatively easily.

program that speeds the entry to a new job presumably increases the likelihood of employment and expected earnings at all future dates. As shown in Table 4, about one-third of the program estimates in our sample, and nearly 60% of the estimates for the Scandinavian countries, are derived from duration models of this form.

b. Summary of Estimated Impacts

As discussed above, we classify each program estimate by whether it is significantly positive, significantly negative, or statistically insignificant. Table 5 presents a tabular summary of the classification rates in our overall sample and in the three country groups. Several features of the data are clear. First, on average the short term impacts are only slightly more likely to be significantly positive (39% of estimates) than significantly negative (28% of estimates). Thus, there appears to be considerable heterogeneity in the measured "success" of ALMP's. Second, the distribution of medium- and long-term outcomes is considerably more favorable than the distribution of short-term outcomes. In the medium term, for example, 50% of the estimated impacts are significantly positive versus 10% significantly negative. The distribution of longer term (3 year) impact estimates is even slightly more favorable, although the sample size is smaller. A third conclusion that emerges from Table 5 is that there are systematic differences across country groups in the distribution of impact estimates. In particular, short-term program impacts appear to be relatively unfavorable in the German-speaking countries, but relatively favorable in the Anglo countries. In the medium term the differences across country groups are smaller, and in the long term the relative position of the German-speaking and Anglo countries is reversed.

Further insights into the relationship between the program impacts at different time horizons is provided in Tables 6a and 6b, which show cross-tabulations between short- and medium-term impacts (Table 6a) or short- and long-term outcomes (Table 6b) for the same program. In both cases, the distributions show a clear tendency for the estimated program impacts to become more favorable over time. For example, none of the programs with an insignificant or significantly positive short-term impact have a significantly negative medium-term impact, but 31% of the programs with a significantly negative short term impact have a significantly positive medium term impact. Likewise, only 2 of the 16 programs that have a significantly negative short-run impact and for which we also have a long-term impact estimate show a significantly negative long-term effect.

Another interesting question is whether there has been any obvious trend in the measured success of active labor market programs. Figure 1a and 1b present some simple evidence suggesting that the answer is "no". The figures show the distributions of short-term and medium term program estimates for programs operated in four time periods: the late 1980s, the early 1990s, the late 1990s, and the post-2000 period. While there is some variability over time, particularly in the distributions of medium term impacts which are based on relatively small samples, there is no tendency for the most recent programs to exhibit better or worse outcomes than programs from the late 1980s.

III. Multivariate Models of the Sign/Significance of Program Estimates

a. Estimating Model

As a motivation for the ordered probit specification we use in our meta analysis, assume

that the i^{th} program estimate, b_i, is derived from a design such that

(1) $$b_i = \beta_i + \frac{k_i \sigma_i}{\sqrt{N_i}} \epsilon_i ,$$

where β_i is the true value of the treatment effect for this program, σ_i represents the standard deviation of the outcome variable used in the evaluation, N_i is the sample size in the evaluation, k_i is a "design effect" (for example, if the evaluation design is a simple comparison of means between a treatment and control group, each of size $N_i/2$, then $k_i = \sqrt{2}$), and ϵ_i is a standard normal variate. This equation represents the sampling error in the i^{th} program estimate as the product of four factors: a design effect, the underlying individual variation in the outcome variable (i.e., σ_i), a sample size factor (the inverse square root of the sample size), and a standard normal variate. In this case the t-statistic associated with the i^{th} program estimate is:

(2) $$t_i = \sqrt{N_i} / k_i \times (\beta_i / \sigma_i) + \epsilon_i$$

where β_i / σ_i is the "effect size" of the i^{th} program treatment. Suppose that in a sample of program estimates $\sqrt{N_i} / k_i$ is approximately constant, and that the effect size in the i^{th} program depends on a set of observable covariates (X_i) and an unobservable component (λ_i):

(3) $$\beta_i / \sigma_i = X_i \alpha' + \lambda_i .$$

Then an appropriate model for the t-statistic is

(4) $$t_i = X_i \alpha + \eta_i$$

where $\alpha = \sqrt{N_i}/k_i \ \alpha'$ and $\eta_i = \epsilon_i + \lambda_i$. Assuming that λ_i is normally distributed, the composite error η_i is normal and equation (4) implies that the probability of observing a t-statistic greater than (or less than) some critical value is given by a probit model that depends on X_i. Likewise, the probability of observing a t-statistic in any ordered partition of the real line is given by an ordered probit model.

The assumption that the "effective sample size" $\sqrt{N_i/k_i}$ is constant across evaluations is surely violated in our sample.[19] Indeed, there is a marginally significant correlation between the probability that the t-statistic for the short-term impact estimate is significant and the square root of the sample size used to obtain the estimate (p-value of the regression coefficient from a linear probability model = 0.03).[20] As explained below we check our inferences by estimating simple probit models for the likelihood of significantly positive or significantly negative program effects that include the square root of the sample size as an additional explanatory variable. In these models we do not see any relationship between the sample size and the probability of a large positive or large negative t-statistic, suggesting that variation in sample size across the studies in our sample is not causing serious biases to our main inferences.

b. Main Estimation Results

Tables 7 and 8 present the main findings from our meta analysis. Table 7 presents a series of models for the likelihood of a significantly positive, significantly negative, or insignificant short-run program estimate, while Table 8 presents a parallel set of models for the medium-term program estimates. We begin by examining the four main dimensions of heterogeneity in our sample separately. Column 1 of each table presents a model that includes a set of dummy variables for the choice of dependent variable used in the study. These are highly

[19]We suspect that the purely mechanical effect of sample size is offset by the tendency of researchers to fit more complex models when they have larger sample sizes, reducing the "effective" sample size (i.e., increasing the design effect associated with the specification). Note too that in many designs the effective sample size is much smaller than the sample size, and depends instead on the number of independent "clusters" in the program and comparison group.

[20]Interestingly, there is no correlation between the square root of the sample size and the probability that the t-statistic for the medium-term estimate is significant (p-value = 0.76).

significant determinants of the measured "success" of a program over a short term horizon. In particular, program estimates derived from hazard models of the time in registered unemployment until exit to a job (row 1), or the time in registered unemployment until any exit (row 2) or the probability of being in registered unemployment (row 4) are more likely to yield a significant positive t-statistic that estimates based on the likelihood of post-program employment (the omitted base group). These patterns are somewhat weaker over the medium term horizon (see Table 8) but still positive, suggesting that evaluations that focus measures of participation in registered unemployment are biased relative to those using other outcomes.[21] We are unsure of the explanation for this finding, although discrepancies between results based on registered unemployment and employment have been noted before in studies of behavior around the point of unemployment benefit exhaustion (see Card, Chetty, and Weber, 2007).[22]

Column 2 of tables 7 and 8 present models that only distinguish the type of program. In the short run, classroom and on-the-job training programs appear to be less successful than the omitted group (combined programs) while job search assistance programs appear (weakly) more successful. Over the medium run, however, the disadvantage of training seems to disappear. Over both horizons, however, subsidized public sector programs are less likely to lead to favorable impact estimates than other types of programs – a finding that emerged in Kluve's (2006) earlier study.

Column 3 presents models that compare program estimates by age and gender. In both

[21]We fit a simplified model using only the 50 program estimates from Scandinavia and found the same pattern. For this subsample, a dummy indicating a register-based outcome measure has a coefficient of 0.61 (standard error=0.34) and is marginally significant.

[22]It is possible for example that assignment to an ALMP causes people to leave benefit system without moving to a job, and that reasons for exit are miscoded.

the short and medium runs a clear finding is that youth programs are relatively unsuccessful. (Again, this echoes one of the main conclusions of Kluve, 2007). It also appears that estimated program effects for participant groups exclude youths are less successful, although we are reluctant to attribute that to a pure age effect, since in most cases the program is open to younger participants but the evaluation only reports results for older participants. We suspect that it may also reflect some unobserved characteristic of the program or the methodology that is shared by the studies that limit attention to older participants. In contrast to the results by age group, the comparisons by gender are never statistically significant. Finally, column 4 presents models that compare shorter and longer duration programs. There is no clear pattern here in either the short- or medium-run impact estimates.

Columns 5 and 6 present models that control for all four dimensions of heterogeneity simultaneously. In Table 7, the specification in column 6 also includes dummies for the type of intake group (people from registered unemployment, long-term unemployed, or disadvantaged workers) dummies for the time period covered by the program (late 1980s, early 1990s, late 1990s, or 2000's) dummies for the three main country groups highlighted in Tables 3 and 4) a dummy for an experimental design, and the square root of the sample size.[23] The parallel specification in Table 8 is more parsimonious (reflecting the smaller sample size available for medium-term program estimates) and includes as extra controls only the experimental design dummy and the square root of the sample size.

Regardless of whether the extra controls are added or not, the pattern of coefficients in

[23]We include the sample size variable for completeness only. Arguably, sample size should affect the probability of a significant negative or positive coefficient. See the discussion in the next subsection.

the multivariate models tends to parallel the patterns in the models that explore one dimension of heterogeneity at a time. In particular, the specifications in columns 5 and 6 of Table 7 suggest that evaluations based on measures of registered unemployment are more likely to show positive impacts in the short run than evaluations based on post-program employment or earnings. Over the longer run this relative advantage appears to weaken (Table 8) though the coefficient estimates from the multivariate model for the medium-term impacts are imprecise. In the short-run, job search assistance programs appear to have a relatively positive impact, while training programs seem to have a bigger advantage in the medium run, and public sector employment programs are uniformly negative. As in the one-dimensional models, programs for youths appear to be relatively unsuccessful in the short or medium runs, while there are no large differences between men and women. Finally, program duration does not seem to matter much either the short or medium runs.

As we noted earlier, many of the studies underlying our sample report separate estimates for male and female participants. This feature allows us to perform a simple but powerful "within-study" comparison of program effectiveness by gender: we simply compare the sign/significance of the program estimates for women and men. For the 28 studies from which we can extract both a short term estimate for women and a short term estimate for men, we found that the estimates were the same (i.e., both significantly positive, both significantly negative, or both insignificant) in 14 cases (50%); the women had a more positive outcome in 8 cases (29%); and the women had a less positive outcome in 6 cases (21%). This comparison provides further evidence that the program outcomes tend to be very similar for women and men (at least at the

crude level of our classification system).[24]

One important finding from the specifications in column 6 of Tables 7 and 8 is that the dummy for experimental evaluations is small in magnitude and statistically insignificant. While the standard errors on the dummy are relatively large, this finding suggests that controlling for outcome measure, program type, and participant group, the non-experimental estimation methods used in the recent literature are yielding roughly the same distributions of sign and significance of the program impacts as the experimental estimators.[25]

Although the coefficients are not reported in Table 7, another notable finding from the specification in column 6 is that the dummies for the country group are jointly insignificant, and all relatively small in magnitude.[26] This suggests that the apparent differences across countries in Table 5 (panel I) are largely explained by differences in the types of program and program participants in the different country groups.

Given the differences between the effects of some of the key covariates on the short term and medium term program impacts, we decided to fit a model for the change in the relative "success" of a given program from the short term to the medium term. To do this, we coded the short term estimate as +1 if the program impact was positive and significant, 0 if the short term

---

[24]A similar conclusion holds when we compare medium term estimates for women and men in the same program: in 8 of the 17 available cases the estimated effects are the same, in 6 cases (35%) women have a more positive outcome and in 3 cases (18%) women have a less positive outcome.

[25]Half of the experimental program estimates use register-based outcome measures. If we include an interaction between experimental design and register-based outcome the coefficient is insignificant (t=0.5), so we do not detect any differential bias in studies that use register-based and other outcome measures, though the power of the test is limited by the small sample of experiments.

[26]The estimated coefficients (and standard errors) are: Germanic countries -0.01 (0.35); Scandinavian countries 0.10 (-0.33); Anglo countries 0.15 (0.53).

impact was insignificant, and -1 if the short term impact was negative and significant. We coded the medium term impact in the same way, and then formed the difference between the measures of medium and short term impacts. The resulting variable takes on values of +2 (for a program that went from negative and significant to positive and significant), +1 (for programs that went from negative and significant to insignificant, or from insignificant to positive and significant), 0 (for programs with the same code in the short and medium runs) and -1 (for the small number of programs that went from significantly positive in the short run to insignificant in the medium run). While the coding system is somewhat arbitrary we believe it captures in a relatively simple way the trend over time in the sign and significance of the impact estimates for any given program.

Ordered probit models fit to the change in impact measure are presented in Table 9. The format follow Tables 7 and 8: the sample size (91 estimates) is the same as in Table 8. The results are somewhat imprecise, but generally confirm the impressions from a simple comparison of the short-term and medium term models. One clear finding is that impact estimates from studies that look at the duration of time in registered unemployment until exit to a job tend to fade between the short term and medium term, relative to impact estimates from other methods (which on average become *more positive*). A second finding is that the impact of training programs tends to rise between the short and medium runs. Interestingly, a similar result has been reported in a recent long term evaluation of welfare reform policies in the U.S. (Hotz, Imbens and Klerman, 2006). This study concludes that although job search assistance programs dominate training in the short run, over longer horizons the gains to human capital development policies are larger.

c. Robustness

One simple way to test the implicit restrictions of our ordered probit model is to fit separate probit models for the events of a significantly positive and significantly negative impact estimate. As noted above, it is also interesting to include a measure of sample size (specifically, the square root of the sample size) in these specifications, because unless researchers are adjusting their designs to hold effective sample size approximately constant, one might expect more large negative t-statistics and more large positive t-statistics from evaluations that use larger samples.

Table 10 shows three specifications for short run program impact. Column 1 reproduces the estimates from the ordered probit specification in column 5 of Table 7. Column 2 presents estimates from a probit model, fit to the event of a significantly positive short run impact. Column 3 presents estimates from a similar probit model, fit to the event of a significantly negative short run impact. Under the assumption that the ordered probit specification is correct, the coefficients in column 2 should be the same as those in column 1, while the coefficients in column 3 should be equal in magnitude and opposite in sign.[27]

Although the coefficients are not in perfect agreement with this prediction, our reading is that the restrictions are qualitatively correct. In particular, the probit coefficients for the covariates that have larger and more precisely estimated coefficients in the ordered probit model (such as the coefficients in rows 1, 2, 7, 9, 10, 11, and 16 of Table 10) fit the predicted pattern very well. Moreover, the coefficients associated with the square root of the sample size (row 18)

[27]A minor issue is that the full set of covariates cannot be included in the model for a significantly negative impact estimate because in some categories of the dummies the dependent variable is 100% predictable.

are relatively small and insignificant in the probit models. This rather surprising finding suggests that variation in sample size is not a major confounding issue for making comparisons across the program estimates in our sample.

d. Estimates for Germany

A concern with any meta analysis that attempts to draw conclusions across studies from many different countries is that the heterogeneity in institutional environments is so great as to render the entire exercise uninformative. Although the absence of large or significant country group effects in our pooled models suggests this may not be a particular problem, we decided to attempt a within-country analysis for the country with the largest number of individual program estimates in our sample, Germany. Since we have only 41 short term impact estimates from Germany, and only 36 medium term estimates, we adopted a relatively parsimonious model that included only 4 main explanatory variables: a dummy for classroom or on-the-job training programs, a dummy for programs with only older (age 25 and over) participants, a measure of program duration (in months), and a dummy for programs operated in the former East Germany.

Results from fitting this specification are presented in Appendix Table A. There are four main findings. First, as in our overall sample, the short run impact of classroom and on-the-job training programs is not much different from other types of programs. But, in the medium run, training programs are associated with significantly more positive impacts. Second, as in our larger sample, it appears that programs for older adults only are less likely to succeed – especially in the medium run – than more broadly targeted programs. Third, longer duration programs are associated with significantly worse short term impacts, but weakly more positive medium term

impacts. This pattern may reflect a mechanical "lock-in" effect of the longer programs, which prevent participants from working in the short run, combined with positive returns to the longer program period. Finally, the models show a negative impact for programs operated in the former East Germany. Overall, we interpret the results from this analysis as quite supportive of the conclusions from our cross-country models.

IV. Summary and Conclusions

Our meta analysis points to a number of important lessons in the most recent generation of active labor market program evaluations. One lesson is that longer-term evaluations are generally more favorable than short-term evaluations. Indeed, we find that many programs that exhibit insignificant or even negative impacts after only a year have significantly positive impact estimates after 2 or 3 years. Classroom and on-the-job training programs appear to be particularly likely to yield more favorable medium-term than short-term impact estimates. A second lesson is that the data source used to measure program impacts matters. Evaluations (including randomized experiments) that measure outcomes based on time in registered unemployment appear to show more positive short-term results than evaluations based on employment or earnings. A third conclusion is that subsidized public sector jobs programs and programs for youth are generally less successful than other types of ALMP's. Here, our findings reinforce the conclusions of earlier literature summaries, including Heckman, Lalonde and Smith (1999), Kluve and Schmidt (2002), and Kluve (2006). A fourth conclusion is that current ALMP programs do not appear to have differential effects on men versus women. Finally, controlling for the program type and composition of the participant group, we find only small and statistically

insignificant differences in the distribution of positive, negative, and insignificant program estimates from experimental and non-experimental evaluations. This is encouraging, and suggests that the research designs used in recent non-experimental evaluations are not significantly biased relative to the benchmark of an experimental design.

Our reading of the literature also points to a number of limitations. Few studies include enough information to perform even a crude cost-benefit analysis (although there are some important counterexamples). For example, program costs are often unknown or unreported. Moreover, the methodological design often precludes making a direct assessment of the program effect on "welfare-relevant" outcomes like earnings, employment, or hours of work. As the methodological issues in the ALMP literature are resolved, we would hope that future studies will adopt a more substantive focus, enabling policy makers to evaluate and compare the social returns to investments in alternative active labor market policies.

References

Ashenfelter, Orley. "Estimating the Effect of Training Programs on Earnings" *Review of Economics and Statistics* 60 (1978): 47-57.

Ashenfelter, Orley. "The Case for Evaluating Training Programs with Randomized Trials." *Economics of Education Review* 6 (1987): 333-338.

Ashenfelter, Orley and David Card. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics* 67 (October 1985): 648-660.

Biewen, Martin, Bernd Fitzenberger, Aderonke Osikominu, and Marie Waller. "Which Program for Whom? Evidence on the Comparative Effectiveness of Public Sponsored Training Programs in Germany." IZA Discussion Paper #2885. Bonn: Institute for the Study of Labor, 2007.

Bring, Johan, and Kenneth Carling. "Attrition and Misclassification of Drop-Outs in the Analysis of Unemployment Duration." *Journal of Official Statistics* 4 (2000): 321–330.

Calmfors, Lars. "Active Labour Market Policy and Unemployment - A Framework for the Analysis of Crucial Design Features, *OECD Economic Studies* 22 (Spring 1994): 7-47.

Card, David, Raj Chetty and Andrea Weber. "The Spike at Benefit Exhaustion: Leaving the Unemployment System or Starting a New Job?" *American Economic Review Papers and Proceedings* 97 (May 2007):113-118.

Card, David and Alan B. Krueger. "Time-Series Minimum-Wage Studies: A Meta-analysis." *American Economic Review Papers and Proceedings* 85 (May 1995): 238-243.

Gerfin Michael and Micael Lechner. "Microeconometric Evaluation of the Active Labour Market Policy in Switzerland." *Economic Journal* 112 (2002): 854-893.

Hagglund, Pathric. "Are There Pre-Programme Effects of Swedish Active Labor Market Policies? Evidence from Three Randomized Experiments." Swedish Institute for Social Research Working Paper 2/2007. Stockholm: Stockholm University, 2007.

Heckman, James J. and Richard Robb. "Alternative Methods for Evaluating the Impact of Interventions." in James J. Heckman and Burton Singer, editors, Longitudinal Analysis of Labor Market Data. Cambridge: Cambridge University Press, 1985: 156-246.

Heckman, James J. , Robert J. Lalonde and Jeffrey A. Smith. "The Economics and Econometrics of Active Labor Market Programs." In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, Volume 3A. Amsterdam and New York: Elsevier, 1999: 1865-2095.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra Todd. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66 (September 1998): 1017-1098.

Heckman, James J., Hiedhiko Ichimura and Petra Todd. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65 (1998): 261-94.

Hedges, Larry V. and Ingram Olkin. *Statistical Methods for Meta-Analysis.* New York: Academic Press, 1985.

Higgins, Julian P.T. and Sally Green (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.0.1. September 2008. Available at www.cochrane-handbook.org.

Hotz, V. Joseph, Guido Imbens and Jacob Klerman. "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Re-Analysis of the California GAIN Program." *Journal of Labor Economics* 24 (July 2006): 521-566.

Jespersen, Svend T., Jakob R. Munch and Lars Skipper. "Costs and Benefits of Danish Active Labour Market Programmes." *Labour Economics* 15 (2008): 859-884.

Johnson, George P. "Evaluating the Macroeconomic Effects of Public Employment Programs. In Orley Ashenfelter and James Blum, editors. *Evaluating the Labor Market Effects of Social Programs.* Princeton, NJ: Princeton University Industrial Relations Section, 1976.

Kluve, Jochen. "The Effectiveness of European ALMP's." In Jochen Kluve et al., *Active Labor Market Policies in Europe: Performance and Perspectives.* Berlin and Heidelberg: Springer, 2007: 153-203.

Lalonde, Robert J. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (September 1986): 604-620.

Richardson Katarina and Gerard J. van den Berg. "The Effect of Vocational Employment Training on the Individual Transition Rate from Unemployment to Work." IFAU Working Paper 2002:8. Institute for Labor Market Policy Evaluation. Uppsala, Sweden, 2002.

Sianesi, Barbara. "An Evaluation of the Swedish System of Active Labor Market Programs in the 199s." *Review of Economics and Statistics* 86 (February 2004): 133-155.

Wilson, Hugh A. "The Development of Americas Postwar Active Labor Market Policy: The Demise of the Two Bang Theory." Unpublished Paper Prepared for Delivery at the Annual Conference of the Midwest Political Science Association, Chicago, April 15 –18, 2004.

Figure 1a: Distribution of Short-term Program Effects Over Time

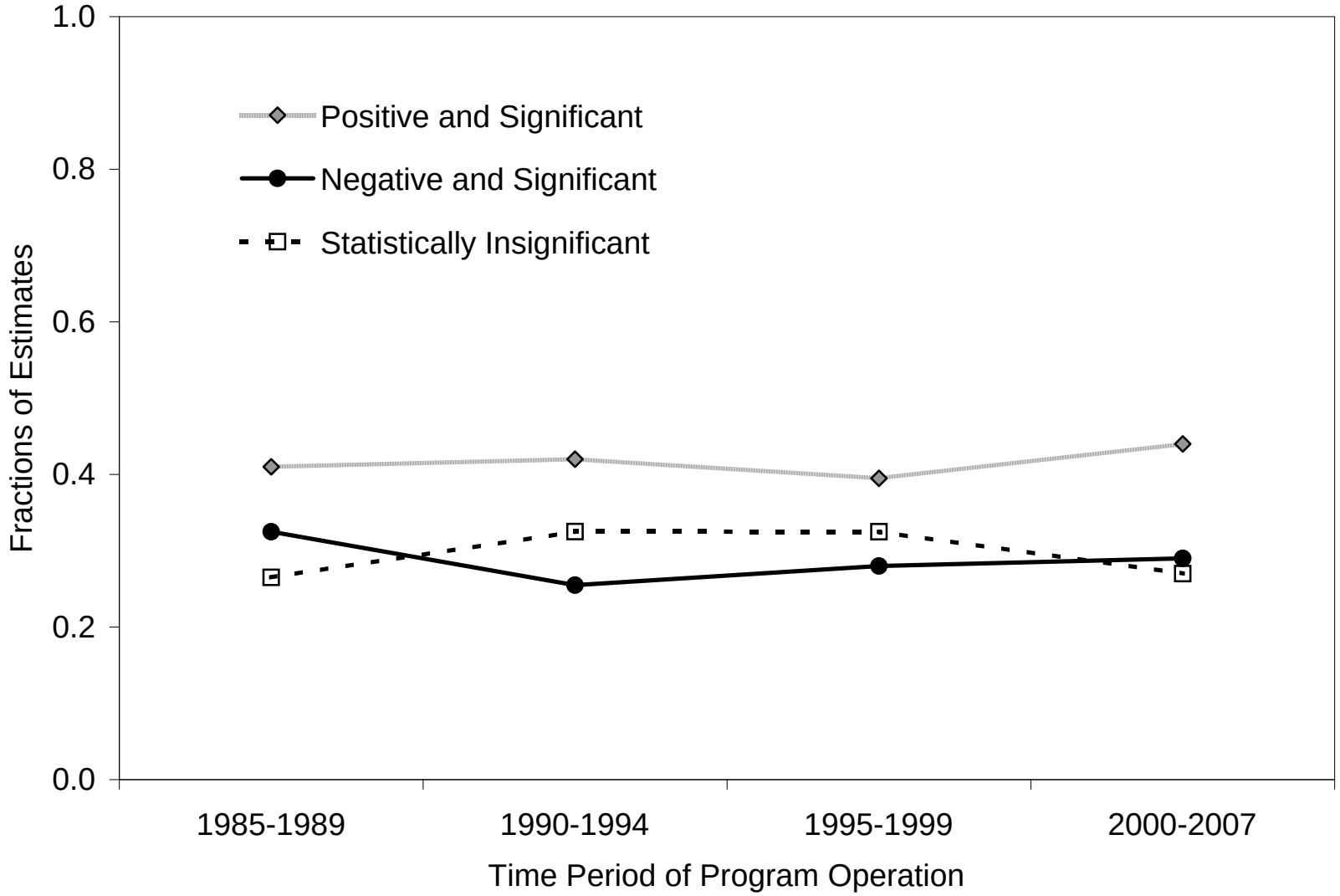

Figure 1b: Distribution of Medium-term Program Effects Over Time

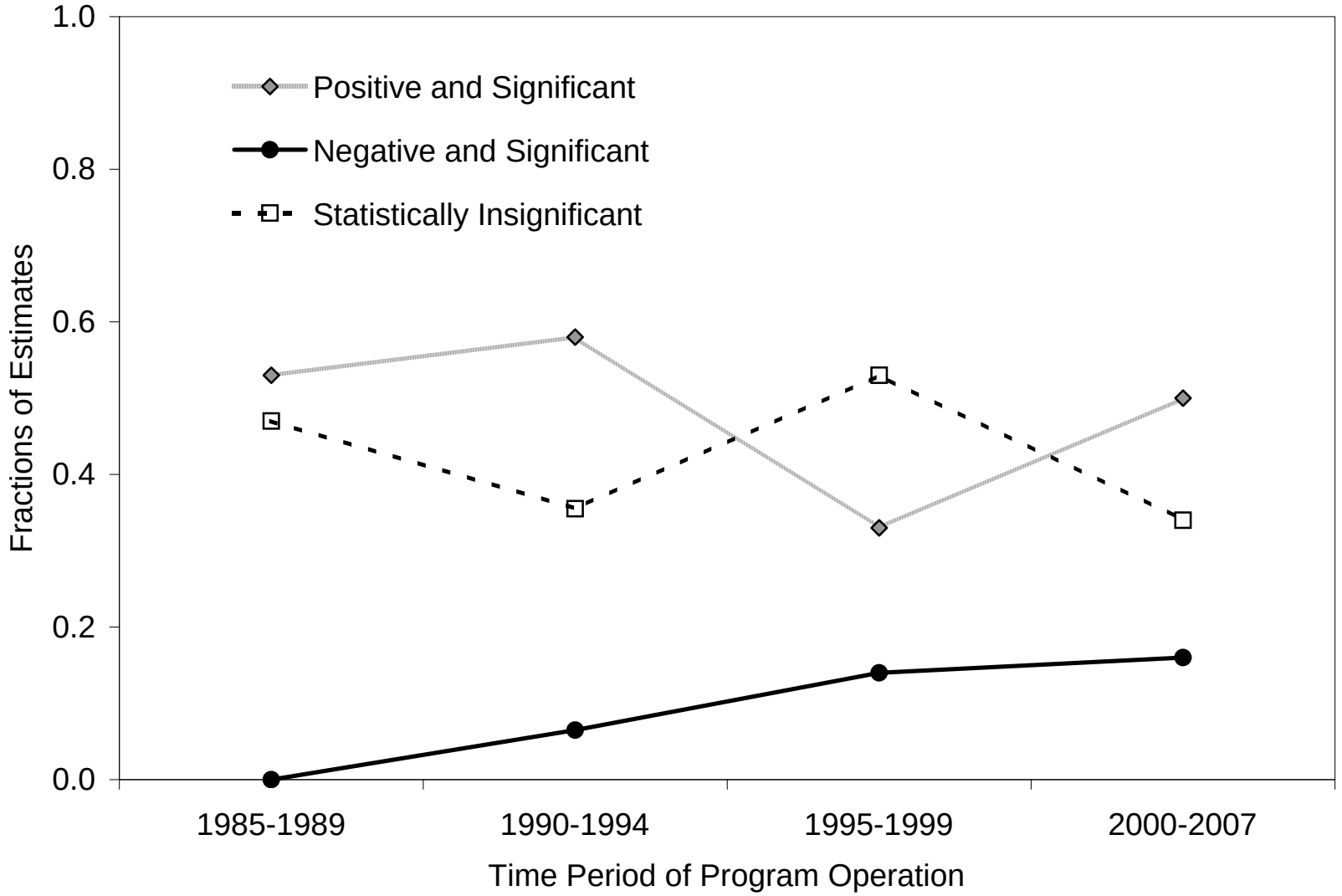

Table 1: Overview of Survey Responses

	Number Contacted (1)	Number Responses (2)	Response Rate (3)	Number with 1+ Research Papers (4)	Percent of Contacts with Papers (5)
1. IZA Fellows	231	152	65.8	66	28.6
2. NBER Labor Studies Associates	130	33	25.4	6	4.6
3. Secondary Contacts	14	12	85.7	12	85.7
4. Total	375	197	52.5	84	22.4

Note: For some sets of co-authors only one co-author responded and their supplied joint paper(s). Thus, the implicit response rate is higher than in column (3). The number of returned papers was 156.

Table 2: Distribution of Program Estimates By Latest Date and Country

	Number of Estimates (1)	Percent of Sample (2)
a. By Latest Date		
1996	2	1.0
1997	2	1.0
1998	4	2.0
1999	12	6.0
2000	10	5.0
2001	3	1.5
2002	19	9.5
2003	14	7.0
2004	26	13.1
2005	16	8.0
2006	41	20.6
2007	48	24.1
2008	2	1.0
b. By Country		
Australia	2	1.0
Austria	13	6.5
Belgium	6	3.0
Canada	1	0.5
Czech Republic	1	0.5
Denmark	25	12.6
Dominican Republic	1	0.5
Estonia	1	0.5
Finland	2	1.0
France	14	7.0
Germany	45	22.6
Hungary	1	0.5
Israel	2	1.0
Netherlands	4	2.0
New Zealand	3	1.5
Norway	7	3.5
Peru	2	1.0
Poland	5	2.5
Portugal	2	1.0
Romania	4	2.0
Slovakia	13	6.5
Spain	3	1.5
Sweden	19	9.5
Switzerland	9	4.5
United Kingdom	4	2.0
United States	10	5.0

Note: Sample includes 199 estimates drawn from 97 separate studies. Date refers to date of the study, whether published or not.

Table 3: Characteristics of Sample of Estimated Program Effects

	Overall Sample (1)	Austria Germany & Switzerland (2)	Scandinavia (3)	Anglo Countries (4)
1. Number of Estimates	199	67	53	20
2. Program Intake				
a. Drawn from Registered Unemployed (%)	68.3	94.0	67.9	15.0
b. Long Term Unemployed (%) (registered and other)	12.6	0.0	3.8	25.0
c. Other (Disadvantaged, etc.) (%)	19.1	6.0	28.3	60.0
3. Type of Program				
a. Classroom or Work Experience Training (%)	41.7	62.7	26.5	35.0
b. Job Search Assistance (%)	12.1	7.5	5.7	30.0
c. Subsidized Private Sector Employment (%)	14.6	3.0	20.8	10.0
d. Subsidized Public Sector Employment (%)	14.1	16.4	9.4	5.0
e. Threat of Assignment to Program (%)	2.5	0.0	7.5	0.0
f. Combination of Types (%)	15.1	10.4	30.2	20.0
4. Program Duration				
a. Unknown or Mixed (%)	26.1	11.9	32.1	45.0
b. 4 Months or Less (%)	20.6	26.9	20.8	25.0
c. 5-9 Months (%)	35.2	28.4	43.4	30.0
d. Over 9 Months (%)	18.1	32.8	3.8	0.0
5. Gender of Program Group[a/]				
a. Mixed (%)	59.3	55.2	73.6	40.0
b. Male Only (%)	20.6	22.1	13.2	25.0
c. Female Only (%)	16.6	21.0	13.2	35.0
6. Age of Program Group[b/]				
a. Mixed (%)	63.8	62.7	56.6	60.0
b. Age Under 25 Only (%)	14.1	0.0	18.9	25.0
c. Age 25 and Older Only (%)	21.6	35.8	24.5	15.0

Notes: Sample includes estimates drawn from 97 separate studies. Scandiavia includes Denmark, Finland, Norway and Sweden. Anglo countries include Australia, Canada, New Zealand, UK, and US.

[a/]When separate estimates are available by gender, a study may contribute estimates for males and females.
[b/]When separate estimates are available by age, a study may contribute estimates for youth and older people.

Table 4: Evaluation Methods Used in Sample of Estimated Program Effects

	Overall Sample (1)	Austria Germany & Switzerland (2)	Scandinavia (3)	Anglo Countries (4)
1. Number of Estimates	199	67	53	20
2. Basic Methodology				
a. Cross Sectional with Comparison Group (%)	3.0	0.0	5.7	0.0
a. Longitudinal with Comparison Group (%)	51.3	80.6	30.2	75.0
c. Duration Model with Comparison Group (%)	36.2	19.4	43.4	0.0
d. Experimental Design (%)	9.1	0.0	18.9	25.0
3. Dependent Variable				
a. Probability of Employment at Future Date (%)	45.7	71.6	17.0	40.0
b. Wage at Future Date (%)	11.6	4.5	20.8	25.0
c. Duration of Time in Registered Unempl. until Exit to Job (%)	24.6	16.4	35.8	10.0
d. Duration of Time in Registered Unempl. (any type of exit) (%)	6.0	1.5	22.6	0.0
e. Other Duration Measures (%)	3.5	0.0	0.0	0.0
f. Probability of Registered Unempl. at Future Date (%)	6.0	6.0	3.8	25.0
4. Covariate Adjustment Method				
a. Matching (%)	50.8	73.1	30.2	45.0
b. Regression (%)	42.7	26.9	52.8	40.0

Notes: See note to Table 1 for definition of country groups.

Table 5: Summary of Estimated Impacts of ALM Programs

	Percent of Estimates that are:		
	Significantly Positive (1)	Insignificant (2)	Significantly Negative (3)
I. Short Term Impact Estimates (~12 Months)			
a. Overall Sample (N=183)	39.3	32.8	27.9
b. Austria, Germany & Switzerland (N=59)	28.8	33.9	37.3
c. Scandinavia (N=50)	46.0	30.0	24.0
d. Anglo Countries (N=17)	70.6	11.8	17.7
II. Medium Term Impact Estimates (~24 Months)			
a. Overall Sample (N=108)	50.0	39.8	10.2
b. Austria, Germany & Switzerland (N=45)	53.3	35.6	11.1
c. Scandinavia (24)	37.5	50.0	12.5
d. Anglo Countries (N=15)	73.3	26.7	0.0
III. Long Term Impact Estimates (36+ Months)			
a. Overall Sample (N=50)	54.0	40.0	6.0
b. Austria, Germany & Switzerland (N=23)	60.9	39.1	0.0
c. Scandinavia (N=15)	40.0	46.7	13.3
d. Anglo Countries (N=10)	50.0	40.0	10.0

Notes: See note to Table 1 for definition of country groups. Significance is based on t-ratio for estimate bigger or smaller than 2.0.

Table 6a: Relation Between Short-Term and Medium-Term Impacts of ALM Programs

	Percent of Medium-Term Estimates that are:		
	Significantly Positive (1)	Insignificant (2)	Significantly Negative (3)
Short Term Impact Estimate:			
a. Significantly Positive (N=30)	90.0	10.0	0.0
b. Insignificant (N=28)	28.6	71.4	0.0
c. Significantly Negative (N=36)	30.6	41.7	27.8

Note: sample includes studies that report short-term and medium-term impact estimates for same program and same participant group.

Table 6b: Relation Between Short-Term and Long-Term Impacts of ALM Programs

	Percent of Long-Term Estimates that are:		
	Significantly Positive (1)	Insignificant (2)	Significantly Negative (3)
Short Term Impact Estimate:			
a. Significantly Positive (N=19)	73.7	21.1	5.3
b. Insignificant (N=13)	30.8	69.2	0.0
c. Significantly Negative (N=16)	43.8	43.8	12.5

Note: sample includes studies that report short-term and long-term impact estimates for same program and same participant group.

Table 7: Ordered Probit Models for Sign/Significance of Estimated Short-term Program Impacts

	Dependent variable = ordinal indicator for sign/significance of estimated impact					
	(1)	(2)	(3)	(4)	(5)	(6)
Dummies for Dependent Variable (omitted=Post-program employment)						
1. Time in Reg. Unemp. Until Exit to Job	0.59 (0.21)	--	--	--	0.45 (0.23)	0.29 (0.26)
2. Time in Registered Unemp.	1.05 (0.33)	--	--	--	1.00 (0.38)	0.99 (0.44)
3. Other Duration Measure	0.38 (0.42)	--	--	--	0.34 (0.44)	0.03 (0.49)
4. Prob. Of Registered Unemp.	1.43 (0.49)	--	--	--	1.37 (0.50)	1.11 (0.53)
5. Post-program Earnings	0.29 (0.30)	--	--	--	0.21 (0.32)	0.03 (0.37)
Dummies for Type of Program (omitted=Mixed and Other)						
6. Classroom or On-the-Job Training	--	-0.40 (0.26)	--	--	-0.04 (0.31)	0.03 (0.36)
7. Job Search Assistance	--	0.38 (0.33)	--	--	0.54 (0.37)	0.65 (0.44)
8. Subsidized Private Sector Job	--	-0.43 (0.31)	--	--	-0.11 (0.34)	-0.12 (0.38)
9. Subsidized Public Sector Job	--	-0.71 (0.32)	--	--	-0.50 (0.37)	-0.46 (0.42)
Dummies for Age and Gender of Participants (omitted=Pooled Age, Pooled Gender)						
10. Age Under 25 Only	--	--	-0.74 (0.25)	--	-0.75 (0.27)	-0.71 (0.30)
11. Age 25 and Older Only	--	--	-0.44 (0.22)	--	-0.40 (0.24)	-0.28 (0.28)
12. Men Only	--	--	-0.11 (0.23)	--	-0.06 (0.24)	-0.16 (0.27)
13. Women Only	--	--	-0.03 (0.22)	--	-0.04 (0.24)	-0.17 (0.27)
Dummies for Program Duration (omitted=5-9 month duration)						
14. Unknown or Mixed	--	--	--	0.46 (0.22)	0.09 (0.26)	0.08 (0.28)
15. Short (≤4 Months)	--	--	--	0.40 (0.22)	0.02 (0.26)	0.11 (0.28)
16. Long (>9 Months)	--	--	--	-0.25 (0.25)	-0.45 (0.28)	-0.44 (0.32)
17. Dummies for Intake Group and Timing of Program	No	No	No	No	No	Yes
18. Dummies for Country Group	No	No	No	No	No	Yes
19. Dummy for Experimental Design	--	--	--	--	--	0.06 (0.39)
20. Square Root of Sample Size (Coefficient × 1000)	--	--	--	--	--	-0.17 (0.27)

Notes: Standard errrors in parentheses. Sample size for all models is 180 program estimates. Models are ordered probit models, fit to ordinal data with value of +1 for significant positive estimate, 0 for insignificant estimate, and -1 for significant negative estimate Estimated cutpoints (2 for each model) are not reported in table.

Table 8: Ordered Probit Models for Sign/Significance of Estimated Medium-term Program Impacts

	Dependent variable = ordinal indicator for sign/significance of estimated impact					
	(1)	(2)	(3)	(4)	(5)	(6)
Dummies for Dependent Variable (omitted=Post-program employment)						
1. Time in Reg. Unemp. Until Exit to Job	0.55	--	--	--	1.21	0.90
	(0.26)				(0.69)	(0.73)
2. Other Duration Measure	0.28	--	--	--	0.38	0.45
	(0.84)				(0.99)	(0.99)
3. Prob. Of Registered Unemp.	0.63	--	--	--	0.33	0.38
	(0.74)				(0.77)	(0.79)
4. Post-program Earnings	0.22	--	--	--	0.04	0.09
	(0.31)				(0.38)	(0.38)
Dummies for Type of Program (omitted=Mixed and Other)						
6. Classroom or On-the-Job Training	--	0.56	--	--	0.86	0.95
		(0.40)			(0.51)	(0.51)
7. Job Search Assistance	--	0.66	--	--	0.48	0.53
		(0.58)			(0.69)	(0.78)
8. Subsidized Private Sector Job	--	0.24	--	--	0.25	0.32
		(0.53)			(0.61)	(0.62)
9. Subsidized Public Sector Job	--	-0.58	--	--	-0.82	-0.80
		(0.47)			(0.60)	(0.60)
Dummies for Age and Gender of Participants (omitted=Pooled Age, Pooled Gender)						
10. Age Under 25 Only	--	--	-0.83	--	-0.89	-0.87
			(0.36)		(0.41)	(0.41)
11. Age 25 and Older Only	--	--	-0.39	--	-1.12	-1.21
			(0.30)		(0.41)	(0.42)
12. Men Only	--	--	-0.40	--	-0.04	-0.17
			(0.34)		(0.40)	(0.42)
13. Women Only	--	--	0.28	--	0.51	0.41
			(0.32)		(0.37)	(0.39)
Dummies for Program Duration (omitted=5-9 month duration)						
14. Unknown or Mixed	--	--	--	-0.72	-1.10	-1.05
				(0.33)	(0.41)	(0.42)
15. Short (≤4 Months)	--	--	--	0.26	-0.43	-0.53
				(0.34)	(0.41)	(0.43)
16. Long (>9 Months)	--	--	--	-0.06	-0.32	-0.28
				(0.33)	(0.39)	(0.39)
17. Dummy for Experimental Design	--	--	--	--	--	0.15
						(0.83)
18. Square Root of Sample Size (Coefficient × 1000)	--	--	--	--	--	1.13
						(0.87)

Notes: Standard errrors in parentheses. Sample size for all models is 91 program estimates. Models are ordered probit models, fit to ordinal data with value of +1 for significant positive estimate, 0 for insignificant estimate, and -1 for significant negative estima Estimated cutpoints (2 for each model) are not reported in table.

Table 9: Ordered Probit Models for Change in Program Impacts from Short-term to Medium-term

	Dependent variable = change in sign/significance of estimated impact					
	(1)	(2)	(3)	(4)	(5)	(6)
Dummies for Dependent Variable (omitted=Post-program employment)						
1. Time in Reg. Unemp. Until Exit to Job	-1.71	--	--	--	-1.36	-1.48
	(0.66)				(0.76)	(0.81)
2. Other Duration Measure	-0.03	--	--	--	0.46	0.48
	(0.81)				(1.00)	(1.00)
3. Prob. Of Registered Unemp.	-1.12	--	--	--	-0.76	-0.64
	(0.82)				(0.90)	(0.92)
4. Post-program Earnings	-0.17	--	--	--	-0.18	-0.16
	(0.32)				(0.38)	(0.39)
Dummies for Type of Program (omitted=Mixed and Other)						
6. Classroom or On-the-Job Training	--	0.88	--	--	0.92	0.96
		(0.44)			(0.57)	(0.58)
7. Job Search Assistance	--	0.10	--	--	0.35	0.54
		(0.62)			(0.75)	(0.82)
8. Subsidized Private Sector Job	--	0.27	--	--	0.24	0.27
		(0.57)			(0.68)	(0.68)
9. Subsidized Public Sector Job	--	0.67	--	--	0.55	0.59
		(0.52)			(0.65)	(0.66)
Dummies for Age and Gender of Participants (omitted=Pooled Age, Pooled Gender)						
10. Age Under 25 Only	--	--	0.29	--	0.20	0.20
			(0.36)		(0.41)	(0.41)
11. Age 25 and Older Only	--	--	-0.01	--	-0.11	-0.11
			(0.29)		(0.39)	(0.40)
12. Men Only	--	--	0.10	--	0.04	-0.02
			(0.35)		(0.41)	(0.43)
13. Women Only	--	--	0.25	--	0.31	0.24
			(0.31)		(0.36)	(0.38)
Dummies for Program Duration (omitted=5-9 month duration)						
14. Unknown or Mixed	--	--	--	-0.92	-0.82	-0.78
				(0.36)	(0.40)	(0.41)
15. Short (≤4 Months)	--	--	--	-0.49	-0.59	-0.57
				(0.34)	(0.40)	(0.41)
16. Long (>9 Months)	--	--	--	0.03	0.09	0.11
				(0.32)	(0.37)	(0.37)
17. Dummy for Experimental Design	--	--	--	--	--	-0.47
						(0.92)
18. Square Root of Sample Size (Coefficient × 1000)	--	--	--	--	--	0.24
						(0.87)

Notes: Standard errrors in parentheses. Sample size for all models is 91 program estimates. Models are ordered probit models, fit to ordinal data with value of +2, +1, 0, and -1, representing the change from the short-term impact (measured as +1, 0 or -1) to the medium-term impact (measured as +1, 0, or -1). Estimated cutpoints (3 for each model) are not reported in table.

Table 10: Comparison of Ordered Probit and Probit Models for Short Term Program Impact

	Ordered Probit (1)	Probit for Significantly Positive Impact (2)	Probit for Significantly Negative Impact (3)
Dummies for Dependent Variable (omitted=Post-program employment)			
1. Time in Reg. Unemp. Until Exit to Job	0.45	0.40	-0.53
	(0.23)	(0.26)	(0.30)
2. Time in Reg. Unemployment	1.10	1.21	-0.76
	(0.42)	(0.45)	(0.58)
3. Other Duration Measure	0.33	-0.53	--
	(0.45)	(0.64)	
4. Prob. Of Registered Unemp.	1.36	1.24	--
	(0.50)	(0.53)	
5. Post-program Earnings	0.20	0.41	0.10
	(0.33)	(0.38)	(0.39)
Dummies for Type of Program (omitted=Mixed and Other)			
6. Classroom or On-the-Job Training	-0.08	0.07	0.14
	(0.34)	(0.39)	(0.51)
7. Job Search Assistance	0.48	0.71	-0.28
	(0.38)	(0.44)	(0.64)
8. Subsidized Private Sector Job	-0.14	0.18	0.37
	(0.36)	(0.43)	(0.54)
9. Subsidized Public Sector Job	-0.54	-0.29	0.69
	(0.39)	(0.46)	(0.56)
Dummies for Age and Gender of Participants (omitted=Pooled Age, Pooled Gender)			
10. Age Under 25 Only	-0.71	-0.85	0.52
	(0.41)	(0.35)	(0.34)
11. Age 25 and Older Only	-0.37	-0.53	0.27
	(0.24)	(0.30)	(0.30)
12. Men Only	-0.02	0.04	0.22
	(0.25)	(0.30)	(0.32)
13. Women Only	0.00	-0.07	-0.02
	(0.24)	(0.29)	(0.31)
Dummies for Program Duration (omitted=5-9 month duration)			
14. Unknown or Mixed	0.07	0.04	-0.18
	(0.26)	(0.31)	(0.34)
15. Short (≤4 Months)	0.06	-0.04	-0.19
	(0.27)	(0.31)	(0.35)
16. Long (>9 Months)	-0.45	-0.32	0.67
	(0.28)	(0.34)	(0.34)
17. Dummy for Experimental Design	0.06	-0.22	--
	(0.36)	(0.41)	
18. Square Root of Sample Size (Coefficient × 1000)	-0.13	0.11	0.31
	(0.22)	(0.24)	(0.29)

Notes: Standard errrors in parentheses. Sample sizes are 180 (cols. 1-2) and 150 (col. 3). Model in column 1 is ordered probit fit to ordinal data with value of +1 for significant positive estimate, 0 for insignificant estimate, and -1 for significantly negative estimate. Model in column 2 is probit for occurrence of signifcantly positive estimate (versus alternative of insigificant or significantly negative). Model in column 3 is probit for occurrence of significantly negative estimate, versus alternative of insigificant or significantly positive estimate.

Appendix Table A: Analysis of Estimated Program Impacts for Germany Only

	Short-term Impact (~12 mo.) (1)	Medium-term Impact (~24 mo.) (2)
Distribution of Dependent Variable:		
%Significant Positive (coded as +1)	24.4	52.8
%Insignificant (coded as 0)	31.7	36.1
%Significant Negative (coded as −1)	43.9	11.1
Coefficients of Ordered Probit Model		
Dummy for Former East Germany (mean=0.44)	-0.24 (0.39)	-1.02 (0.58)
Dummy for Classroom or On-the-job Training (mean=0.80)	0.17 (0.54)	3.13 (0.93)
Dummy for Participants Age 25 or Older Only (mean=0.46)	-0.77 (0.42)	-1.19 (0.98)
Program Duration in Months (mean=8.91)	-0.09 (0.04)	0.05 (0.06)
Number of Estimates	41	36

Notes: standard errors in parentheses. Models also include a dummy for observations with imputed value for program duration. Estimated cut-points for ordered probit (2 for each model) are not reported.