THE
ECONOMIC
JOURNAL

The Economic Journal, **120** (*November*), F452–F477. Doi: 10.1111/j.1468-0297.2010.02387.x. © 2010 The Author(s). The Economic Journal © 2010 Royal Economic Society. Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

ACTIVE LABOUR MARKET POLICY EVALUATIONS: A META-ANALYSIS*

David Card, Jochen Kluve and Andrea Weber

This article presents a meta-analysis of recent microeconometric evaluations of active labour market policies. We categorise 199 programme impacts from 97 studies conducted between 1995 and 2007. Job search assistance programmes yield relatively favourable programme impacts, whereas public sector employment programmes are less effective. Training programmes are associated with positive medium-term impacts, although in the short term they often appear ineffective. We also find that the outcome variable used to measure programme impact matters, but neither the publication status of a study nor the use of a randomised design is related to the sign or significance of the programme estimate.

The effectiveness of active labour market policies – including subsidised 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 labour market programme (ALMP) affects the labour market outcomes of the participants themselves.[2] Progress has been facilitated by rapid advances in methodology and data quality, 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 *et al.* (1999) summarise approximately 75 microeconometric evaluation studies from the US and other countries. A more recent review by Kluve (forthcoming) includes nearly 100 separate studies from Europe alone, while Greenberg *et al.* (2003) survey 31 evaluations of government-funded programmes for the disadvantaged in the US.

In this article we synthesise some of the main lessons in the recent microeconometric evaluation literature, using a new and comprehensive sample of programme estimates from the latest generation of studies. Our sample is derived from responses to a survey of 358 academic researchers affiliated with the Institute for the Study of Labour (IZA) and the National Bureau of Economic Research (NBER) in spring 2007. These researchers and their colleagues authored a total of 97 studies of active labour market

* Corresponding author: David Card, University of California Berkeley. Email: card@econ.berkeley.edu.

We thank the authors who responded to our survey for their co-operation and assistance. We also thank four anonymous referees for helpful suggestions, and participants of an IZA/World Bank workshop, participants of LAMES 2009 Buenos Aires, and seminar participants at the University of Chile and the Inter-America Development Bank for comments on earlier versions of the article. Our research was funded by the Center for Labor Economics at UC Berkeley, the German Science Foundation (DRG, SFB475), and the Leibniz Association (Pakt für Foschung und Innovation).

[1] In the US, for example, the direct public sector employment programmes initiated by the Works Progress Administration in 1935 were immediately controversial.

[2] A key unsettled question is whether ALMPs 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 programmes and Calmfors (1994) for a more recent critique, focusing on the European experience of the 1980s and early 1990s.

policies between 1995 and 2007 that meet our inclusion criteria.[3] We conduct a meta-analysis using a sample of 199 'programme estimates' – estimated effects for a particular programme on a specific group of participants – 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 programme – and a medium-term estimate giving the effect approximately 2 years after completion. We also have longer-term (3 year) impacts for one-quarter of the programmes. These estimates allow us to compare shorter and longer-term effects of different types of programmes, and test whether certain programme features are associated with either a larger or smaller programme impact in the short run than in the longer run.

In our main analysis we classify the estimates by whether the post-programme impact on the participants is found to be significantly positive, statistically insignificant or significantly negative. This simple classification of sign and significance allows us to draw comparisons across studies that use very different dependent variables – ranging from the duration of time in registered unemployment to average quarterly earnings – and very different econometric modelling strategies. As a check we also examine the estimated 'effect sizes' from the largest subgroup of studies that focus on participants' employment probabilities and compare meta-analytic models for the estimated effect size with those that use only the sign and significance of the programme effects. We find that the two approaches yield very similar conclusions about the role of programme type, participant characteristics and the evaluation methodology on the measured effectiveness of active labour market programmes.

Consistent with earlier summaries, our analysis suggests that subsidised public sector employment programmes are relatively ineffective, whereas job search assistance (JSA) and related programmes have generally favourable impacts, especially in the short run. Classroom and on-the-job training programmes are not particularly effective in the short run, but have more positive relative impacts after two years. Comparing across different participant groups, we find that programmes for youths are less likely to yield positive impacts than untargeted programmes, although in contrast to some earlier reviews we find no large or systematic differences by gender (Bergemann and van den Berg, forthcoming). We also find that evaluations based on the duration of time in registered unemployment are more likely to show positive short-term impacts than those based on direct labour market outcomes (i.e. employment or earnings).

An important issue in the ALMP evaluation literature is the difficulty of controlling for selection biases that may lead to specious positive or negative programme effects.[4] This concern led observers in the 1980s to call for randomised programme evaluations (e.g., Ashenfelter, 1987). In recent years a significant number of randomised trials have been conducted, and randomised designs account for nearly 10% of the estimates in our sample. This feature allows us to compare the results of experimental and non-experimental evaluations, while controlling for the nature of the programme and

[3] Of these 97 studies, 37 were included in the evaluation by Kluve (forthcoming). Most of the others are very recent – see below.

[4] See, e.g., Ashenfelter (1978), Ashenfelter and Card (1985), Heckman and Robb (1985), Lalonde (1986), Heckman, Ichimura *et al.* (1998), and Heckman *et al.* (1999). Imbens and Wooldridge (2009) present a survey of the most recent methodological advances in programme evaluation.

its participants. We find that the mean differences between the experimental and non-experimental impact estimates are small and statistically insignificant ($t < 0.5$). We also test for potential 'publication bias' (Easterbrook *et al.*, 1991) by examining whether published studies are more or less likely to find a positive effect but find no indication of such a pattern.

The next Section of the article describes our sample of recent microeconometric evaluation studies and the criteria we used for including a study in our analysis sample. Section 2 presents a descriptive overview of the programme estimates we extracted from the included studies. Section 3 presents our main meta-analysis results. Section 4 concludes the article.

1. Assembling a New Sample of ALMP Programme Estimates

1.1. *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 Labour (IZA) and the National Bureau of Economic Research (NBER).[5] We obtained the email list for IZA research fellows who had indicated an interest in the programme area 'Evaluation of labour market programmes', and the list of associates of the NBER Labour Studies programme. 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 programme 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.[6]

Our list of IZA fellows was extracted on January 25, 2007 and contained a total of 231 names and valid email addresses (excluding the three of us). We emailed the survey on February 21st, 2007. We followed a similar procedure for affiliates of the NBER Labour Studies Programme, extracting names and email addresses on March 20, 2007, and emailing the survey to 113 NBER affiliates who were not on the IZA list 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 that constitute a third subgroup in our sample.

Table 1 summarises the responses to our survey. The overall response rate across the 358 researchers we ultimately contacted was 55%. 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.[7] Among the 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.

[5] 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 64% of the programme estimates in our sample are derived from published studies.

[6] The questionnaire is available on request.

[7] The response rate for the 17 NBER members who are also part of IZA was 47%.

Table 1

Overview of Survey Responses

	Number Contacted	Number Responses	Response Rate	Number with 1+ Papers	% of Contacts with Papers
	(1)	(2)	(3)	(4)	(5)
1. IZA Fellows	231	152	65.8	66	28.6
2. NBER Labor Studies Affiliates	113	33	29.2	6	5.3
3. Secondary Contacts	14	12	85.7	12	85.7
4. Total	358	197	55.0	84	23.5

Note. Results from survey of IZA members with interest in 'Evaluation of Labour Market Programmes', conducted January 2007, and of NBER Labour Studies affiliates, conducted March 2007.
Secondary contacts were referred by original sample members.

1.2. *Selection of Studies*

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

- classroom or on-the-job training
- job search assistance or sanctions for failing to search[8]
- subsidised private sector employment
- subsidised public sector employment

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

Methodologically, we decided to limit our attention to well-documented empirical evaluation studies based on individual micro data. We also excluded a few studies that lacked an explicit comparison group of people who were not subject to the programme (or who entered the programme at a later date).

Applying these rules, we eliminated 33 of the originally submitted studies that did not meet our ALMP programme requirements and 18 that did not meet our methodological criteria. We also eliminated 8 studies that were written in a language other than English,[9] or had substantial overlap with other studies included in the sample (e.g.,

[8] A couple of programmes are actually based on the *threat* of assignment to a programme, which we interpret as a form of sanction: see e.g., Hagglund (2007). Sanctions, threats and JSA programmes are all short-term programmes with little (or no) 'lock-in' or 'incapacitation' effect – so participants can enter the labour market very soon after entering the programme.

[9] We included studies in other languages if the author(s) returned a completed questionnaire.

earlier versions of the same study), or were otherwise incomplete. The remaining 97 studies (=156−33−18−8) form the basis for our empirical analysis. A complete list of the studies included in our analysis sample is contained in the online Appendix.

1.3. *Extraction of Programme Estimates and Other Information*

The third step in our data collection process was to extract information about the programme and participants analysed in each study and the estimated programme impact(s). Although we initially intended to collect these data from the questionnaires distributed in our email survey, we were unable to do so because only 38% of authors returned a questionnaire (and many of these were only partially complete). Ultimately, we decided to extract the information ourselves.[10]

Some variables were relatively straightforward to collect, including the type of programme, the age and gender of the participant population, the type of dependent variable used to measure the impact of the programme 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, for example, relatively few studies present detailed information on the pre-programme characteristics of the participants and the comparison group.[11] Another (surprising) fact is that very few studies provide information on programme costs. We decided to use average programme duration as a rough proxy for the size of the investment represented by the programme.

The most difficult task, however, proved to be the development of a standardised measure of programme impact that could be compared across studies. This is mainly due to the wide variation in methodological approaches in the literature. For example, about one-third of the studies in our sample report treatment effects on the exit rate from registered unemployment. Very rarely do these studies include enough information to infer the cumulated effect of the programme on the probability of employment at some date after the completion of the programme.

Faced with such a diverse set of outcome measures and modelling strategies we abandoned the preferred meta-analytic approach of extracting a standardised 'effect size' estimate from each study.[12] Instead, we classified the estimates based on 'sign and significance' into three categories: significantly positive, insignificantly different from zero and significantly negative.[13] Whenever possible, we extracted the sign and significance of the programme impact at three points: a *short-term* impact at approxi-

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

[11] See e.g., Heckman, Ichimura *et al.* (1998) and Heckman, Ichimura and Todd (1998).

[12] The effect size is usually defined as the ratio of the treatment effect on the treated population to the standard deviation of the outcome variable. See Hedges and Olkin (1985).

[13] 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. Vote counting is problematic when individual studies have low power (so an insignificant outcome is likely, even when the true effect is non-zero).

mately one year after completion of the programme, a *medium-term* impact roughly two years after programme completion and a *long-term impact* roughly three years after programme completion.

While we were unable to extract standardised effect sizes from our full sample of studies, for a subset of 35 studies that measure programme effects on the probability of employment we were able to extract an estimated programme effect and the associated employment rate of the comparison group. For these studies we define the estimated 'effect size' as the ratio of the estimated programme effect to the standard deviation of employment among the comparison group. In Section 3, below, we compare meta-analytic models fit to the programme estimates from this subset of studies using our 'sign and significance' measure and the estimated effect size of the programme.

Many studies in our sample report separate impacts for different programme types (e.g., job training versus private sector employment) and/or for different participant subgroups. Whenever possible, we extracted separate estimates for each programme 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 'programme estimates' (estimates for a specific programme and participant group) from the 97 studies in our sample.[14] For many of the programme/subgroup combinations we have a short-term impact estimate and a medium and/or long-term impact. Specifically, for 54% of the programme/subgroup combinations we have a short-term and medium-term programme impact, while for 24% we have a short-term and a long-term impact estimate.

1.4. *Sample Overview*

Table 2 shows the distribution of our sample of programme estimates by the latest publication date of the study (panel *a*) and by country (panel *b*).[15] The studies included in our sample are all relatively recent: 90% of the programme estimates come from articles or working papers dated 2000 or later and 50% from papers dated 2006 or later. Just under two-thirds of the estimates are taken from published studies (measuring publication status as of January 2010). The estimates cover a total of 26 countries, with the largest numbers from Germany (45 estimates), Denmark (26 estimates), Sweden (19 estimates) and France (14 estimates).

2. Descriptive Analysis

2.1. *Programme Types, Participant Characteristics, and Evaluation Methodology*

Table 3 presents a summary of the programme types and participant characteristics represented in our sample of 199 programme estimates. To facilitate discussion we have defined three broad 'country groups' that together account for about 70% of the programme estimates. Countries in each group share many important institutional

[14] A total of 56 studies contribute a single programme estimate, 17 studies contribute 2 estimates, and 24 studies contribute 3 or more estimates.

[15] Note that 46% of the estimates are from unpublished studies. By 'publication date' we mean the date on the study, whether published or not.

Table 2

Distribution of Programme Estimates By Latest Date and Country

	Number of Estimates	% of Sample
	(1)	(2)
(*a*) *By Latest Revison or Publication Date:*		
1996	2	1.0
1997	2	1.0
1998	3	1.5
1999	13	6.5
2000	10	5.0
2001	4	2.0
2002	18	9.1
2003	13	6.5
2004	20	10.1
2005	12	6.0
2006	29	14.6
2007	39	19.6
2008	14	7.0
2009	13	6.5
2010	7	3.5
Published Studies	128	64.3
(*b*) *By Country of Programme*		
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

Notes. Sample includes 199 estimates from 97 separate studies.

features and also tend to have similar design features in their active labour market programmes. The largest group of estimates is from Austria, Germany and Switzerland (AGS) with 67 programme estimates (column 2 of Table 3). The second largest group is from the Nordic countries (Denmark, Finland, Norway and Sweden) with 53 programme estimates (column 3). A third distinct group is the ‘Anglo’ countries (Australia,

Table 3

Characteristics of Sample of Estimated Programme Effects

	Overall Sample	Austria Germany & Switzerland	Nordic Countries	Anglo Countries
	(1)	(2)	(3)	(4)
1. Number of Estimates	199	67	53	20
2. *Programme 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 Programme*				
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. Subsidised Private Sector Employment (%)	14.6	3.0	20.8	10.0
d. Subsidised Public Sector Employment (%)	14.1	16.4	9.4	5.0
e. Threat of Assignment to Programme (%)	2.5	0.0	7.5	0.0
f. Combination of Types (%)	15.1	10.4	30.2	20.0
4. *Programme 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 Programme Group**				
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 Programme Group*†				
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. Nordic Countries include Denmark, Finland, Norway and Sweden. Anglo countries include Australia, Canada, New Zealand, UK and US.
*When separate estimates are available by gender, a study may contribute estimates for males and females.
†When separate estimates are available by age, a study may contribute estimates for youth and older people.

Canada, New Zealand, UK and US). For this group – summarised in column 4 of Table 3 – we have 20 programme estimates.

The entries in rows 2*a*–2*c* of Table 3 illustrate one of the most important dimensions of heterogeneity between the three main country groups, which is the intake source of ALMP participants. In Austria, Germany and Switzerland, most active labour market programmes are provided to people in registered unemployment and participation is generally mandatory. Some 94% of the programme estimates for AGS are for such programmes. In the Anglo countries, by comparison, many programmes are targeted to long-term disadvantaged individuals who voluntarily enroll though community outreach programmes. Nearly 60% of the programme estimates for these countries are for these types of participants. The Nordic countries are closer to AGS: about two-thirds of programme estimates are for programmes provided to the registered unemployed and just under one-third are for other disadvantaged groups.

The entries in rows 3*a*–3*f* show the types of active labour market programmes in our sample. Classroom and work experience training programmes are the most common, particularly in AGS, where 63% of the programme estimates are for classroom or on-the-job training programmes. Job search assistance and sanction programmes are relatively uncommon in AGS and the Nordic countries but are more widespread in the Anglo countries.[16] Subsidised public and private employment programmes together account for about 30% of our sample of programme estimates and are relatively evenly distributed across the three main country groups. Finally, combination programmes are particularly common in the Nordic countries, where people who remain in registered unemployment often are automatically assigned to some form of 'active' programme (Sianesi, 2004).

Rows 4*a*–4*d* show the distribution of programme durations. In general, most active labour market programmes are short, with a typical duration of 4–6 months. Programmes tend to be somewhat longer in AGS and shorter in the Anglo countries. The short duration of the programmes 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. Given the modest investment (and opportunity cost) of a 4–6 month programme, an impact on the order of a 5–10% permanent increase in labour market earnings might be large enough to justify the programme on a cost-benefit basis.[17]

Rows 5*a*–*c* and 6*a*–*c* of Table 3 present data on the gender and age composition of the participant groups associated with the programme estimates. Our reading of the programme descriptions leads us to believe that few of the programmes 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 programmes* on men and women. The situation with respect to age is somewhat different. Sometimes the programmes are specifically targeted on younger workers (i.e., those under 21 or 25), whereas sometimes programmes are available to all age groups but the analysts have limited their study to participants over the age of 24, or stratified by age.[18] In any case, most of the programme estimates in our sample are for pooled age and gender groups.

Table 4 describes the features of the evaluation methods used in our sample. Apart from the randomised designs, there are two main methodological approaches in the recent literature. One, which is widely adopted in AGS and the Anglo countries, uses longitudinal administrative data on employment and/or earnings for the participants and a comparison group (who are assigned to a simulated starting date for a potential programme). Typically, the data set includes several years of pre-programme labour market history and propensity-score matching is used to narrow the comparison group

[16] 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 programmes included in our sample are special programmes outside these usual services (or in some cases provided to people who are not in registered unemployment).

[17] Jespersen *et al.* (2008) present a detailed cost-benefit analysis for various Danish programmes and conclude that subsidised public and private sector employment programmes have a positive net social benefit, whereas classroom training programmes do not.

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

Table 4

Evaluation Methods Used in Sample of Estimated Programme Effects

	Overall Sample	Austria Germany & Switzerland	Nordic Countries	Anglo Countries
	(1)	(2)	(3)	(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
b. 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 Unemployment until Exit to Job (%)	24.6	16.4	35.8	10.0
d. Duration of Time in Registered Unemployment (any type of exit) (%)	8.5	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 3 for definition of country groups.

to a sample whose observed characteristics and pre-programme outcomes closely match those of the participants (Gerfin and Lechner, 2002; Biewen *et al.*, 2007; Jespersen *et al.*, 2007). In this type of study, the programme effect is usually measured in terms of the probability of employment at some date after the completion of the programme, although earnings can also be used. Over two-thirds of the evaluations from AGS and the Anglo countries fit this mould, as do a minority (about 30%) of the evaluations from the Nordic countries.

The main alternative approach, widely used in the Nordic countries, is a duration model of the time to exit from registered unemployment – see e.g., Sianesi (2004). The popularity of this approach is due in part to the fact that in many countries all the necessary data can be drawn from the benefit system itself (i.e., without having access to employment records). The programme effect is parameterised as the difference in the exit rate from registered unemployment between participants who entered a specific programme at a certain date and the exit rate of a comparison group who did not. In some studies the outcome variable is defined as the exit rate *to a new job* while in others the exit event includes all causes.[19] Even in the former case, however, the programme effect cannot be easily translated into an impact on employment rates.[20] Nevertheless,

[19] Bring and Carling (2000) show that in the Swedish case nearly one-half of those who exit for other reasons are later found to be working, so the classification by reason for exit is noisy.

[20] Richardson and Van den Berg (2002) show that with a constant programme entry rate and a proportional effect on the hazard to employment the effect on employment can be derived.

the sign of the treatment effect is interpretable, since a programme 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 programme estimates in our sample, and nearly 60% of the estimates for the Nordic countries, are derived from models of this form.

2.2. *Summary of Estimated Impacts*

As discussed above, our main analysis focuses on the 'sign and significance' of the programme estimates. Table 5 presents a tabular summary of programme estimates in our overall sample and the three broad country groups, classified by whether the estimate is significantly positive, insignificant, or significantly negative. The entries in row 1*a* show that on average the short-term impacts (measured roughly one year after programme completion) are slightly more likely to be significantly positive (39% of estimates) than significantly negative (25% of estimates). Thus, there appears to be considerable heterogeneity in the measured 'success' of ALMPs. Second, the distribution of medium and long-term outcomes is more favourable than the distribution of short-term outcomes. In the medium term, for example, 45% of the estimated impacts

Table 5
Summary of Estimated Impacts of ALM Programmes

	Percent of Estimates that are:		
	Significantly Positive	Insignificant	Significantly Negative
	(1)	(2)	(3)
1. *Short-term Impact Estimates* (*∼12 Months After Completion of Programme*)			
a. Overall Sample (*N* = 184)	39.1	36.4	24.5
b. Austria, Germany & Switzerland (*N* = 59)	28.8	40.7	30.5
c. Nordic Countries (*N* = 50)	46.0	30.0	24.0
d. Anglo Countries (*N* = 18)	66.7	16.7	16.6
e. Outcome Measure = Probability of Employment (*N* = 79)	25.3	41.7	33.0
f. Median Effect Size for Estimates with Outcome = Probability of Employment (*N* = 76)	0.21	0.01	−0.21
2. *Medium-term Impact Estimates* (*∼24 Months After Completion of Programme*)			
a. Overall Sample (*N* = 108)	45.4	44.4	10.2
b. Austria, Germany & Switzerland (*N* = 45)	44.4	44.4	11.1
c. Nordic Countries (*N* = 24)	37.5	50.0	12.5
d. Anglo Countries (*N* = 15)	66.7	33.3	0.0
e. Outcome Measure = Probability of Employment (*N* = 66)	39.4	47.0	13.6
f. Median Effect Size for Estimates with Outcome = Probability of Employment (*N* = 59)	0.29	0.03	−0.20
3. *Long-term Impact Estimates* (*36 + Months After Completion of Programme*)			
a. Overall Sample (*N* = 51)	52.9	41.1	6.0
b. Austria, Germany & Switzerland (*N* = 23)	60.9	39.1	0.0
c. Nordic Countries (*N* = 15)	40.0	46.7	13.3
d. Anglo Countries (*N* = 11)	45.5	45.5	9.0

Notes. See note to Table 3 for definition of country groups. Significance is based on t-ratio for estimate bigger or smaller than 2.0. Effect size for observations with outcome measure = probablity of employment equals estimated treatment effect divided by standard deviation of outcome in the control group.

are significantly positive versus 10% significantly negative. The distribution of longer-term (3 years after programme completion) impact estimates is even more favourable, although the sample size is smaller.

A third interesting conclusion from Table 5 is that there are systematic differences across country groups in the distribution of impact estimates. In particular, short-term impacts appear to be relatively unfavourable in Austria, Germany and Switzerland but relatively favourable in the Anglo countries. One explanation for this pattern is the heterogeneity across country groups in the types of programmes. In fact, as we discuss below, once we control for the type of programme and other features, the cross-country differences narrow and are no longer significant.

As mentioned earlier, we extracted standardised 'effect size' estimates for a subsample of evaluations that use the post-programme probability of employment as the outcome of interest. In this subsample the fraction of significantly positive short-term estimates is slightly lower than in the sample as a whole, while the fraction of significantly negative estimates is slightly higher (compare row 1*e* with row 1*a*). The medium-term impacts, however, have about the same distribution as in the overall sample (compare row 2*e* with row 2*a*). Row 1*f* of Table 5 shows the average short-term effect sizes for employment-based studies in each of the three categories, while row 2*f* shows the average medium-term effect sizes among studies in each category.[21] As might be expected, the mean effect size for the 'insignificant' programme estimates is very close to 0. More surprising, perhaps, is that the mean effect size for significantly positive short-term estimates (0.21) is equal in magnitude but opposite in sign to the mean effect size for significantly negative short-term estimates (−0.21). This symmetry is consistent with the assumption that the t-statistic for a programme estimate is proportional to the effect size. As we discuss below, in this case an analysis of the sign and significance of the programme estimates yields the same conclusions as an analysis of the effect size of different programmes.

The relationship between the programme impacts at different time horizons is illustrated in Tables 6*a* and 6*b*, which show cross-tabulations between short and medium-term impacts (Table 6*a*) or short and long-term outcomes (Table 6*b*) *for the same programme*. In both cases the estimated programme impacts appear to become more positive over time. For example, 31% of the programmes with a significantly negative short-term impact have a significantly positive medium-term impact, whereas none of the programmes with an insignificant or significantly positive short-term impact has a significantly negative medium-term impact. Likewise, most of the programmes with a significantly negative short-term impact show either a significantly positive or insignificant long-term impact.

One important question in the evaluation literature is whether ALMPs have become more effective over time; see e.g., the discussion in Lechner and Wunsch (2006). Figures 1*a* and 1*b* present some simple evidence suggesting that the answer is 'no'. The Figures show the distributions of short-term and medium-term programme estimates for programmes 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

[21] Recall that the effect size is the estimated programme effect in the probability of employment, divided by the average employment rate of the control group.

Table 6*a*

Relation Between Short-term and Medium-term Impacts of ALM Programmes

	% of Medium-term Estimates that are:		
	Significantly Positive	Insignificant	Significantly Negative
	(1)	(2)	(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 programme and same participant group.

Table 6*b*

Relation Between Short-term and Long-term Impacts of ALM Programmes

	% of Long-term Estimates that are:		
	Significantly Positive	Insignificant	Significantly Negative
	(1)	(2)	(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 programme and same participant group.

the distributions of medium term impacts, which are based on relatively small samples, there is no tendency for the most recent programmes to exhibit better or worse outcomes than programmes from the late 1980s.

3. Multivariate Models of the Sign/Significance of Programme Estimates

3.1. *Meta-Analytic Model*

We begin by discussing the conditions under which an analysis of the sign and significance of the programme estimates from a sample of studies is informative about the actual effectiveness of the underlying programmes. Assume that the *ith* programme estimate, b, is derived from an econometric procedure such that b is normally distributed around the true treatment effect β with variance V^2/N, where N represents the overall sample size used in the evaluation, i.e.,

$$b \sim \mathrm{N}(\beta, V^2/N).$$

Assume that $V = K\sigma$, where σ is the standard deviation of the outcome variable used in the evaluation (e.g., $\sigma =$ standard deviation of earnings per month) and K

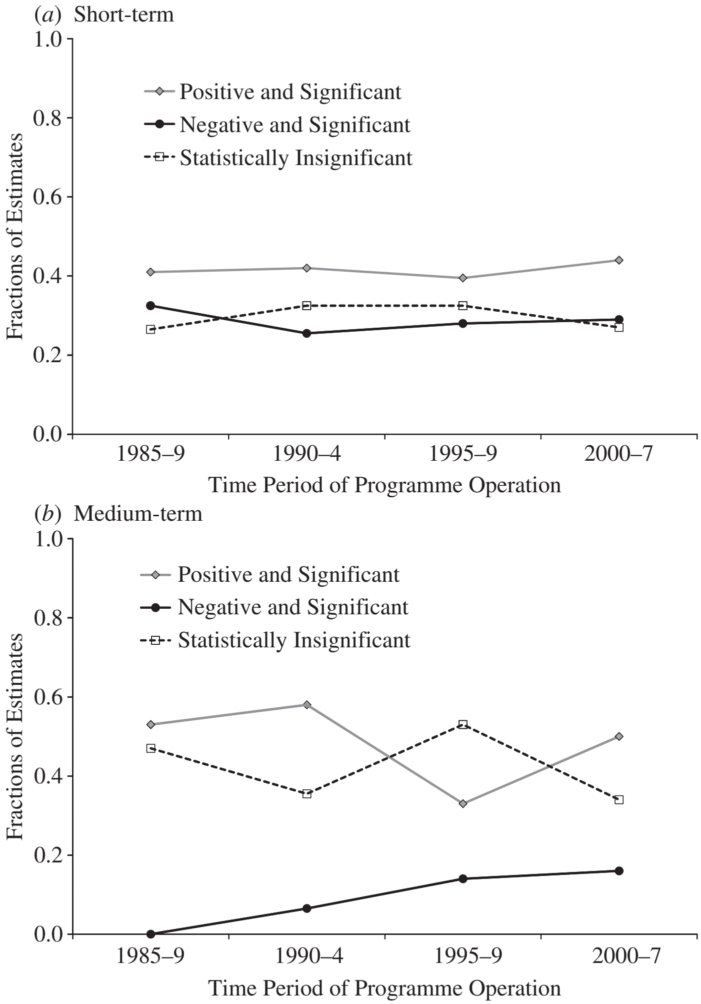

Fig. 1. *Distribution of Programme Effects Over Time*

is a study design factor. The realised value of the programme estimate b can be written as

$$b = \beta + N^{-1/2}K\sigma z \tag{1}$$

where z is the realisation of a standard normal variate. The 't-statistic' associated with the estimated treatment effect is

$$\mathrm{t} = b/\mathrm{Var}(b)^{1/2} = (N^{1/2}/K) \times (\beta/\sigma) + z. \tag{2}$$

Note that β/σ is the 'effect size' of the programme. Equation (2) implies that the observed t-statistic differs from a realisation of a standard normal variate by a term that reflects a combination of the effect size of the programme, the sample size, and the design effect K.[22]

Suppose that in a sample of programme estimates the ratio $N^{½}/K$ is constant, and that the effect size of the *ith* programme depends on a set of observable covariates (**X**):

$$\beta/\sigma = \mathbf{X}\boldsymbol{\alpha}. \tag{3}$$

Under these assumptions an appropriate model for the observed t-statistic from the *i*th programme is

$$\mathrm{t} = \mathbf{X}\boldsymbol{\alpha}' + z, \tag{4}$$

where $\boldsymbol{\alpha}' = (N^{½}/K)\ \boldsymbol{\alpha}$. Since z is normally distributed (with variance 1) (4) implies that the probability of observing a significantly negative programme estimate ($\mathrm{t} \leq -2$), an insignificant estimate ($-2 < \mathrm{t} < 2$) or a significantly positive estimate ($\mathrm{t} \geq 2$) is given by an ordered probit model with index function $\mathbf{X}\boldsymbol{\alpha}'$.

The value of $N^{½}/K$ – which can be interpreted as the 'effective sample size' underlying a given programme estimate, taking account of sample size and design complexity – clearly varies across the evaluations included in our sample. Nevertheless, we believe that the tendency for researchers to use more complex research designs (with higher values of K) when bigger samples are available tends to largely offset the 'mechanical' effect of sample size on the distribution of t-statistics. In this case an analysis of the 'sign and significance' of the programme estimates provides a useful guide to the determinants of the effectiveness of ALMPs.

We conduct two specification tests designed to evaluate the validity of our main analysis, based on sign and significance of the programme estimates. First, we estimate simple probit models for the likelihood of significantly positive or significantly negative programme effects that include the square root of the sample size as an additional explanatory variable. If the 'effective' sample size systematically varies with the actual sample size (i.e., if $N^{½}/K$ is increasing in N) then we would expect to see more large positive t-statistics *and* more large negative t-statistics from studies with larger sample sizes. In fact, as we show below, these simple probit models show *no relationship* between the sample size and the probability of either a significantly positive or significantly negative t-statistic, confirming that the 'mechanical' effect of sample size is mitigated by other design factors.

As a second test, we fit an alternative meta-analysis model to the programme estimates derived from the subsample of programmes that measure impacts of the probability of employment. For these studies we extract an estimate of the effect size b/s, where s is the estimated standard deviation of the outcome of interest (employment status) in the comparison group. Assuming that

$$b/s = \beta/\sigma + \varepsilon,$$

where ε represents the sampling error for the estimated effect size in the *i*th programme, (3) implies that:

[22] For example in a randomised evaluation with equal-sized treatment and control groups, if the programme causes a simple shift in the mean of the treatment group then $K = 2$.

$$b/s = \mathbf{X}\boldsymbol{\alpha} + \varepsilon. \tag{5}$$

We therefore fit a linear regression of the estimated effect sizes on the covariates **X** and compare the vector of estimated coefficients to the estimates from our ordered probit specification. If $N^{1/2}/K$ is indeed constant, then the coefficients from the ordered probit model of sign and significance and the OLS model of effect sizes should be proportional (with a factor of proportionality $= \boldsymbol{\alpha}'/\boldsymbol{\alpha} = N^{1/2}/K$).

3.2. *Main Estimation Results*

Tables 7 and 8 present the main findings from our meta-analysis. Table 7 shows a series of models for the likelihood of a significantly positive, significantly negative, or insignificant short-run programme estimate, while Table 8 presents a parallel set of models for the medium-term programme estimates, and for the change between the short-term and medium-term estimates. We begin in Table 7 by separately examining the four main dimensions of heterogeneity across the studies in our sample. The model in column 1 includes a set of dummy variables for the choice of outcome variable used in the study. These are highly significant determinants of the short-term 'success' of a programme (i.e., roughly one year after programme completion). In particular, programme estimates derived from 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 than those derived from models of post-programme employment (the omitted base group). We are unsure of the explanation for this finding, although discrepancies between results based on registered unemployment and employment have been noted before in the literature (Card *et al.*, 2007).[23]

The model in column 2 of Table 7 summarises the patterns of sign and significance for different programme types. In the short run, classroom and on-the-job training programmes appear to be less successful than the omitted group (combined programmes) while job search assistance programmes appear (weakly) more successful. The 'least successful' programmes are subsidised public sector jobs programmes – a result that parallels the findings in Kluve's (forthcoming) study of an earlier group of studies.

The model in column 3 compares programme estimates by age and gender. Interestingly, the programme estimates for people under 25 and those age 25 and over *both* appear to be more negative than the estimates for mixed age groups. We suspect this pattern reflects some combination of programme characteristics and other factors that are shared by the studies that estimate separate effects by age (rather than an effect of participant age *per se*). In contrast to the results by age, the comparisons by gender are never statistically significant.[24] Finally, column 4 presents models that compare shorter

[23] It is possible for example that assignment to an ALMP causes people to leave the benefit system without moving to a job. In this case programmes will appear to be more effective in reducing registered unemployment than in increasing employment.

[24] We were able to extract separate short-term programme estimates for men and women in the same programme from a total of 28 studies. Within this subgroup, the estimates for the two gender groups have the same sign/significance in 14 cases (50%); the women have a *more* positive outcome in 8 cases (29%); and the women have a *less* positive outcome in 6 cases (21%). The symmetry of these comparisons provides further evidence that programme outcomes tend to be very similar for women and men.

Table 7

Ordered Probit Models for Sign/Significance of Estimated Short-term Programme Impacts

	Dependent variable = ordinal indicator for sign/significance of estimated impact					
	(1)	(2)	(3)	(4)	(5)	(6)
Dummies for Dependent Variable (*omitted = Post-programme employment*)						
1. Time in Reg. Unemp. Until Exit to Job	0.47	–	–	–	0.34	0.18
	(0.20)				(0.24)	(0.28)
2. Time in Registered Unemp.	0.85	–	–	–	0.84	0.88
	(0.36)				(0.39)	(0.49)
3. Other Duration Measure	0.29	–	–	–	0.17	−0.07
	(0.21)				(0.31)	(0.31)
4. Prob. Of Registered Unemp.	1.38	–	–	–	1.22	0.92
	(0.47)				(0.58)	(0.66)
5. Post-programme Earnings	0.26	–	–	–	0.09	−0.07
	(0.37)				(0.38)	(0.48)
Dummies for Type of Programme (*omitted = Mixed and Other*)						
6. Classroom or On-the-Job Training	–	−0.30	–	–	0.04	0.22
		(0.26)			(0.30)	(0.38)
7. Job Search Assistance	–	0.35	–	–	0.41	0.72
		(0.34)			(0.36)	(0.44)
8. Subsidised Private Sector Job	–	−0.50	–	–	−0.25	−0.14
		(0.31)			(0.35)	(0.42)
9. Subsidised Public Sector Job	–	−0.67	–	–	−0.50	−0.31
		(0.38)			(0.37)	(0.46)
Dummies for Age and Gender of Participants (*omitted = Pooled Age, Pooled Gender*)						
10. Age Under 25 Only	–	–	−0.70	–	−0.67	−0.69
			(0.29)		(0.28)	(0.32)
11. Age 25 and Older Only	–	–	−0.55	–	−0.57	−0.51
			(0.25)		(0.27)	(0.30)
12. Men Only	–	–	−0.10	–	−0.03	−0.11
			(0.24)		(0.22)	(0.24)
13. Women Only	–	–	−0.03	–	0.00	−0.07
			(0.23)		(0.21)	(0.25)
Dummies for Programme Duration (*omitted = 5–9 month duration*)						
14. Unknown or Mixed	–	–	–	0.42	0.07	0.10
				(0.24)	(0.25)	(0.28)
15. Short ($\leq$ 4 Months)	–	–	–	0.33	−0.04	0.00
				(0.22)	(0.27)	(0.29)
16. Long (>9 Months)	–	–	–	−0.07	−0.24	−0.24
				(0.31)	(0.35)	(0.38)
17. Dummies for Intake Group and Timing of Programme	No	No	No	No	No	Yes
18. Dummies for Country Group	No	No	No	No	No	Yes
19. Dummy for Experimental Design	–	–	–	–	–	−0.06
						(0.40)
20. Square Root of Sample Size (Coefficient × 1000)	–	–	–	–	–	−0.02
						(0.03)
21. Dummy for Published	–	–	–	–	–	−0.18
						(0.26)
Pseudo R-squared	0.04	0.03	0.03	0.02	0.10	0.12

Notes. Standard errrors (clustered by study) in parentheses. Sample size for all models is 181 programme estimates. Models are ordered probits, 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 the Table.

Table 8

Ordered Probit Models for Sign/Significance of Medium-term Impacts and Change in Impact from Short-term to Medium-term

	Medium-term Impact		Change in Impact: Short-term to Medium-term	
	(1)	(2)	(3)	(4)
Dummies for Dependent Variable (omitted = Post-programme employment)				
1. Time in Reg. Unemp. Until Exit to Job	1.29	0.95	−0.10	0.66
	(0.68)	(1.04)	(1.01)	(1.30)
2. Other Duration Measure	0.63	0.21	1.07	2.40
	(0.46)	(1.05)	(0.43)	(1.30)
3. Prob. Of Registered Unemp.	0.59	0.15	−0.55	−0.97
	(0.93)	(1.04)	(0.32)	(0.51)
4. Post-programme Earnings	0.45	0.65	0.18	−0.11
	(0.34)	(0.63)	(0.32)	(0.57)
Dummies for Type of Programme (omitted = Mixed and Other)				
5. Classroom or On-the-Job Training	0.74	1.14	0.81	0.84
	(0.49)	(0.68)	(0.34)	(0.64)
6. Job Search Assistance	0.49	1.16	0.38	0.42
	(0.61)	(0.85)	(0.40)	(0.88)
7. Subsidised Private Sector Job	0.36	0.79	0.22	0.38
	(0.62)	(0.92)	(0.58)	(0.65)
8. Subsidised Public Sector Job	−0.92	−0.46	0.40	0.24
	(0.57)	(0.74)	(0.43)	(0.68)
Dummies for Age and Gender of Participants (omitted = Pooled Age, Pooled Gender)				
9. Age Under 25 Only	−0.82	−0.96	0.15	0.79
	(0.28)	(0.53)	(0.30)	(0.55)
10. Age 25 and Older Only	−0.92	−0.83	−0.12	−0.16
	(0.41)	(0.52)	(0.44)	(0.60)
11. Men Only	0.03	−0.28	0.31	0.47
	(0.32)	(0.45)	(0.31)	(0.45)
12. Women Only	0.32	0.17	0.34	0.29
	(0.36)	(0.44)	(0.28)	(0.44)
Dummies for Programme Duration (omitted = 5–9 month duration)				
13. Unknown or Mixed	−1.08	−1.57	−0.89	−1.00
	(0.33)	(0.46)	(0.37)	(0.46)
14. Short (≤ 4 Months)	−0.29	−0.41	−0.61	−0.35
	(0.36)	(0.46)	(0.44)	(0.52)
15. Long (>9 Months)	−0.34	−0.50	−0.30	−0.36
	(0.30)	(0.37)	(0.50)	(0.68)
16. Dummy for Experimental Design		0.41	–	−0.12
		(0.83)		(0.70)
17. Square Root of Sample Size (Coefficient × 1000)		0.13	–	−0.10
		(0.13)		(0.11)
18. Dummy for Published		−0.08		0.61
		(0.34)		(0.33)
Pseudo R-squared	0.19	0.26	0.09	0.14

Notes. Standard errrors (clustered by study) in parentheses. Sample size for all models is 92 programme estimates. Models in columns 1 and 2 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. Models in columns 3 and 4 are ordered probit models, fit to ordinal data with values of +2, +1, 0, and −1, representing the change form the short-term impact to the medium-term impact. Estimated cutpoints are not reported in the Table.

and longer duration programmes. There is no evidence here that longer duration programmes are more effective than short programmes.

Columns 5 and 6 of Table 7 present models that control for all four dimensions of heterogeneity simultaneously. The extended specification in column 6 also includes dummies for the intake group (registered unemployed, disadvantaged workers, or long-term unemployed), the time period of the programme (in 5-year intervals), and the three main country groups, as well as controls for experimental design, sample size, and publication status. As in the simpler models, the coefficients from the multivariate models suggest that evaluations based on measures of registered unemployment are more likely to show positive short-term impacts than those based on post-programme employment or earnings, while job search assistance programmes have more positive impacts than training or subsidised employment programmes. The gender and age effects in columns 5 and 6 are similar to those from the model in column 3, and the programme duration effects are not too different from the effects in the model in column 4.

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.[25] Thus, differences in the outcome variable, the type of programme and the characteristics of programme participants appear to explain the rather large differences across countries that are apparent in rows 1*b*–1*d* of Table 5.

The estimated coefficients in column 6 associated with experimental designs, published studies and sample size are all small in magnitude and statistically insignificant. The estimate of the experimental design effect suggests that controlling for the outcome measure, the programme type and the participant group, non-experimental estimation methods tend to yield the same distribution of sign and significance as experimental estimators.[26] Likewise the estimated coefficient for published studies suggests that these are no more (or less) likely to show significantly positive programme effects than their unpublished counterparts. The sample size effect is harder to interpret, since if larger samples lead to more precise results we might expect offsetting effects on the likelihood of obtaining significantly positive and significantly negative effects. We return to the sample size effect in the 'one-sided' probit models below.

Columns 1 and 2 of Table 8 present a parallel set of ordered probit models for the medium-term programme effects (measured about 2 years after programme completion). Given the smaller number of medium-term programme estimates (91 versus 180 short-term estimates) the extended specification in column 2 of Table 8 includes controls for sample size, experimental design and publication status but excludes the controls for intake group, time period and country group. Although the standard errors are relatively large, the estimates point to several notable differences in the determinants of short-term and medium term impacts.

[25] The estimated coefficients (and standard errors) are: AGS 0.06 (0.35); Nordic countries 0.23 (0.34); Anglo countries 0.05 (0.55), all relative to the omitted group of other countries.

[26] Half of the experimental programme estimates use register-based outcomes. If we include an interaction between experimental design and register-based outcome the coefficient is insignificant (t = 0.5), so there is no indication of a differential bias in studies that use register-based and other outcome measures, though the power of the test is limited.

To evaluate these differences more carefully, we decided to fit a set of models for the *change* in the relative 'success' of a given programme from the short term to the medium term. Specifically we coded the change as +2 if the programme estimate changed from significantly negative in the short term to significantly positive in the medium term, +1 if the estimate moved from significantly negative to insignificant, or from insignificant to significantly positive, 0 if the short-term and medium-term estimates were classified the same, and −1 if the estimate moved from significantly positive to insignificant, or from insignificant to significantly negative. While the coding system is somewhat arbitrary we believe it captures the trend over time in the sign and significance of the impact estimates for any given programme.

Ordered probit models fitted to the change in impact measure are presented in columns 3 and 4 of 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 probability of registered unemployment 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 programmes 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 US (Hotz *et al.*, 2006). This study concludes that although job search assistance programmes dominate training in the short run, over longer horizons the gains to human capital development policies are larger.

3.3. *Evaluating the Meta-Analysis Model*

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 9 shows three specifications for short-run programme 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 2, 7, 9, 10, and 11 of

[27] The full set of covariates cannot be included in the probit model for a significantly negative impact estimate because some covariates predict the outcome perfectly.

Table 9

Comparison of Ordered Probit and Probit Models for Short-term Programme Impact

	Ordered Probit	Probit for Significantly Positive Impact	Probit for Significantly Negative Impact
	(1)	(2)	(3)
Dummies for Dependent Variable (omitted = Post-programme employment)			
1. Time in Reg. Unemp. Until Exit to Job	0.32	0.39	−0.24
	(0.24)	(0.27)	(0.32)
2. Time in Reg. Unemployment	0.94	0.99	−0.86
	(0.46)	(0.49)	(0.67)
3. Other Duration Measure	0.17	−0.64	–
	(0.32)	(0.52)	
4. Prob. Of Registered Unemp.	1.21	1.11	–
	(0.58)	(0.59)	
5. Post-program Earnings	0.10	0.36	0.22
	(0.38)	(0.38)	(0.42)
Dummies for Type of Programme (omitted = Mixed and Other)			
6. Classroom or On-the-Job Training	0.06	0.08	−0.04
	(0.36)	(0.38)	(0.56)
7. Job Search Assistance	0.42	0.53	−0.42
	(0.38)	(0.44)	(0.65)
8. Subsidised Private Sector Job	−0.21	0.01	0.41
	(0.40)	(0.46)	(0.59)
9. Subsidised Public Sector Job	−0.44	−0.31	0.60
	(0.45)	(0.48)	(0.62)
Dummies for Age and Gender of Participants (omitted = Pooled Age, Pooled Gender)			
10. Age Under 25 Only	−0.68	−0.89	0.50
	(0.29)	(0.34)	(0.36)
11. Age 25 and Older Only	−0.56	−0.80	0.38
	(0.27)	(0.29)	(0.33)
12. Men Only	−0.01	0.02	0.17
	(0.23)	(0.25)	(0.28)
13. Women Only	0.01	−0.08	−0.10
	(0.21)	(0.26)	(0.27)
Dummies for Programme Duration (omitted = 5–9 month duration)			
14. Unknown or Mixed	0.09	0.14	−0.06
	(0.27)	(0.29)	(0.38)
15. Short ($\leq$4 Months)	−0.03	0.07	0.21
	(0.29)	(0.33)	(0.43)
16. Long (>9 Months)	−0.22	−0.01	0.46
	(0.35)	(0.39)	(0.41)
17. Dummy for Experimental Design	0.05	−0.26	–
	(0.32)	(0.41)	
18. Square Root of Sample Size (Coefficient × 1000)	−0.01	0.02	0.03
	(0.02)	(0.02)	(0.03)
19. Dummy for Published	−0.13	0.02	0.30
	(0.19)	(0.23)	(0.26)
Pseudo R-squared	0.09	0.15	0.11

Notes. Standard errrors in parentheses. Sample sizes are 181 (cols. 1–2) and 150 (col. 3). Model in column 1 is ordered probit fit to ordinal data with value of +1 for significantly positive estimate, 0 for insignificant estimate, and −1 for significantly negative estimate. Model in column 2 is probit model for event of significantly positive effect. Model in column 3 is probit for event of significantly negative estimate.

Table 9) fit the predicted pattern very well. Moreover, the coefficients associated with the square root of the sample size (row 18) 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 programme estimates in our sample.

Our second specification test compares the results from an ordered probit model based on sign and significance of the programme estimates to a simple linear regression fit to the estimated effect sizes from different programmes. For this analysis we use a sample of 79 programme estimates (derived from 34 studies) that use the probability of employment as an outcome. A series of specifications for the two alternative meta-analytic models is presented in Table 10. Columns 1–4 present a set of ordered probit

Table 10

Comparison of Models for Sign/Significance and Effect Size of Short-term Programme Estimates, Based on Subsample of Studies with Probability of Employment as Dependent Variable

	Ordered Probit Models for Sign / Significance:				OLS Regressions for Effect Size:			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dummies for Type of Programme (omitted = Mixed and Other)								
1. Classroom or On-the-Job Training	−0.98	–	–	−0.89	−0.23	–	–	−0.14
	(0.52)			(0.54)	(0.12)			(0.12)
2. Job Search Assistance	−0.41	–	–	−0.83	−0.07	–	–	−0.11
	(0.66)			(0.69)	(0.08)			(0.18)
3. Subsidised Private Sector Job	−1.50	–	–	−1.41	−0.35	–	–	−0.24
	(0.62)			(0.60)	(0.12)			(0.13)
4. Subsidised Public Sector Job	−2.46	–	–	−2.54	−0.46	–	–	−0.38
	(0.60)			(0.67)	(0.11)			(0.12)
Dummies for Age and Gender of Participants (omitted = Pooled Age, Pooled Gender)								
5. Age Under 25 Only	–	−1.11	–	−0.97	–	−0.32	–	−0.26
		(0.47)		(0.59)		(0.04)		(0.04)
6. Age 25 and Older Only	–	−0.72	–	−1.14	–	−0.19	–	−0.23
		(0.44)		(0.39)		(0.08)		(0.06)
7. Men Only	–	−0.73	–	−0.51	–	−0.16	–	−0.11
		(0.49)		(0.36)		(0.09)		(0.07)
8. Women Only	–	−0.08	–	0.12	–	−0.07	–	−0.04
		(0.50)		(0.30)		(0.08)		(0.05)
Dummies for Programme Duration (omitted = 5–9 month duration)								
9. Unknown or Mixed	–	–	0.29	−0.37	–	–	0.09	−0.04
			(0.39)	(0.52)			(0.13)	(0.09)
10. Short (≤ 4 Months)	–	–	0.59	0.08	–	–	0.06	−0.07
			(0.34)	(0.53)			(0.05)	(0.08)
11. Long (>9 Months)	–	–	0.04	−0.60	–	–	−0.04	−0.15
			(0.44)	(0.53)			(0.08)	(0.07)
Pseudo R-squared / R-squared	0.13	0.09	0.02	0.23	0.23	0.28	0.03	0.46

Notes. Standard errrors (clustered by study) in parentheses. Sample sizes are 79 (col 1–4) and 76 (col 5–8). Only progamme estimates from studies that use the probability of employment as the outcome are included. Models in columns 1–4 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 estimates. Models in columns 5–8 are linear regression models, fit to the effect size defined by the programme impact on the treatment group over average outcome in the control group. Estimated cutpoints from ordered probit models are not reported in the Table.

models that parallel the models in columns 2–5 of Table 7. Estimates from the subset of studies that use the probability of employment as an outcome measure are generally similar to the estimates from the wider sample: in particular, subsidised public sector programmes appear to be relatively ineffective, while programme estimates for participants under age 25 and for age 25 and older both tend to be relatively unfavourable. More importantly, the estimates from the ordered probit models in columns 1–4 appear to be very close to linear rescalings of the coefficients from the OLS models in columns 5–8 (with a scale factor of roughly 6), as would be predicted if the t-statistics for the programme estimates are proportional to the associated effect sizes. This consistency is illustrated in Figure 2, where we plot the ordered probit coefficients from the model in column 4 of Table 10 against the corresponding OLS coefficients from the model in column 8. The two sets of estimates are very highly correlated ($\rho = 0.93$) and lie on a line with slope of roughly 6.5 and intercept close to 0. (The t-statistic for the test that the intercept is 0 is 1.32).

Overall, we interpret the estimates in Table 10 and the pattern of coefficients in Figure 2 as providing relatively strong support for the hypothesis that the t-statistics associated with the programme estimates in the recent ALMP evaluation literature are proportional to the underlying effect sizes. Under this assumption, a vote-counting analysis (i.e., a probit analysis for the event of a significantly positive estimate), an ordered probit analysis of sign and significance, and a regression analysis of the effect size all yield the same conclusions about the determinants of programme success. Surprisingly, perhaps, this prediction is confirmed by the models in Tables 9 and 10.

3.4. *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

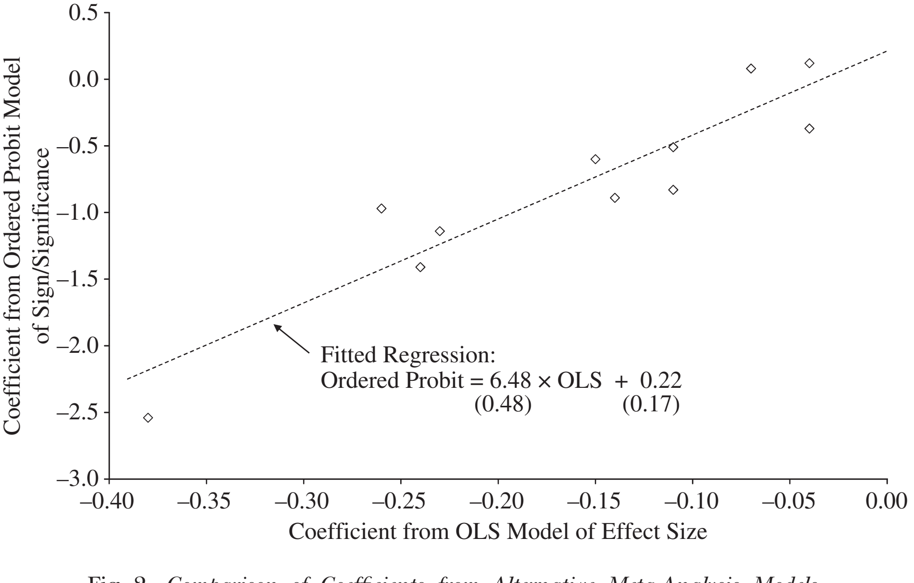

Fig. 2. *Comparison of Coefficients from Alternative Meta-Analysis Models*

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 programme 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 programmes, a dummy for programmes with only older (age 25 and over) participants, a measure of programme duration (in months), and a dummy for programmes operated in the former East Germany.

Results from fitting this specification are presented in Appendix Table A1. There are four main findings. First, as in our overall sample, the short-run impact of classroom and on-the-job training programmes is not much different from other types of programmes. But, in the medium run, training programmes are associated with significantly more positive impacts. Second, as in our larger sample, it appears that programmes for older adults only are less likely to succeed – especially in the medium run – than more broadly targeted programmes. Third, longer duration programmes are associated with significantly worse short-term impacts but weakly more positive medium-term impacts. Finally, the models show a negative impact for programmes 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.

4. Summary and Conclusions

Our meta-analysis points to a number of important lessons in the most recent generation of active labour market programme evaluations. One is that longer-term evaluations tend to be more favourable than short-term evaluations. Indeed, it appears that many programmes with 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 programmes appear to be particularly likely to yield more favourable medium-term than short-term impact estimates. A second lesson is that the data source used to measure programme impacts matters. Evaluations (including randomised 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 subsidised public sector jobs programmes are generally less successful than other types of ALMPs. Here, our findings reinforce the conclusions of earlier literature summaries, including Heckman *et al.* (1999), Kluve and Schmidt (2002), and Kluve (forthcoming). A fourth conclusion is that current ALMP programmes do *not* appear to have differential effects on men versus women. Finally, controlling for the programme type and composition of the participant group, we find only small and statistically insignificant differences in the distribution of positive, negative and insignificant programme estimates from experimental and non-experimental evaluations, and between published and unpublished studies. The absence of an 'experimental' effect suggests that the research designs used in recent non-experimental evaluations are not significantly biased relative to the benchmark of an experimental design. The similarity between published and unpublished studies likewise eases concern over the potential for 'publication bias'.

Methodologically, our analysis points to a potentially surprising feature of the recent generation of programme estimates, which is that the t-statistics from the programme estimates appear to be (roughly) proportional to the effect sizes, and independent of the underlying sample sizes used in the evaluation. In this case a simple 'vote-counting analysis' of significantly positive effects, or an ordered probit analysis of sign and significance, yield the same conclusions about the determinants of programme success as a more conventional meta-analytic model of programme effect sizes. We conjecture that researchers tend to adopt more sophisticated research designs when larger sample sizes are available, offsetting the purely mechanical impact of sample size on the t-statistics that is emphasised in much of the meta-analysis literature.

Our reading of the ALMP literature also points to a number of limitations of the most recent generation of studies. Most importantly, few studies include enough information to make even a crude assessment of the benefits of the programme relative to its costs. Indeed, many studies completely ignore the 'cost' side of the evaluation problem. Moreover, the methodological designs adopted in the literature often preclude a direct assessment of the programme effect on 'welfare-relevant' outcomes like earnings, employment, or hours of work. As the methodological issues in the ALMP literature are resolved, we anticipate that future studies will adopt a more substantive focus, enabling policy makers to evaluate and compare the social returns to investments in alternative active labour market policies.

University of California Berkeley
RWI – Essen
University of Mannheim and RWI-Essen

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix A: Analysis of Estimated Programme Impacts for Germany Only

Appendix B: List of Studies Included in Analysis Sample

Please note: The RES and Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing materials) should be directed to the authors of the article.

References

Ashenfelter, O. (1978). 'Estimating the effect of training programs on earnings', *Review of Economics and Statistics*, vol. 60, pp. 47–57.

Ashenfelter, O. (1987). 'The case for evaluating training programs with randomized trials', *Economics of Education Review*, vol. 6, pp. 333–8.

Ashenfelter, O. and Card, D. (1985). 'Using the longitudinal structure of earnings to estimate the effect of training programs', *Review of Economics and Statistics*, vol. 67 (October), pp. 648–60.

Bergemann, A. and van den Berg, G. (forthcoming). 'Active labour market policy effects for women in Europe: a survey', *Annales d'Economie et de Statistique*.

Biewen, M., Fitzenberger, B., Osikominu, A. and Waller, M. (2007). 'Which program for whom? Evidence on the comparative effectiveness of public sponsored training programs in Germany', IZA Discussion Paper No. 2885, Bonn: Institute for the Study of Labour.

Bring, J. and Carling, K. (2000). 'Attrition and misclassification of drop-outs in the analysis of unemployment duration', *Journal of Official Statistics*, vol. 4, pp. 321–30.

Calmfors, L. (1994). 'Active labour market policy and unemployment – a framework for the analysis of crucial design features', *OECD Economic Studies*, vol. 22, pp. 7–47.

Card, D, Chetty, R. and Weber, A. (2007). 'The spike at benefit exhaustion: leaving the unemployment system or starting a new job?', *American Economic Review Papers and Proceedings*, vol. 97, pp. 113–8.

Easterbrook, P.J., Berlin, A., Gopalan, R. and Matthews, D.R. (1991). 'Publication bias in clinical research', *Lancet*, vol. 337, pp. 867–72.

Gerfin, M. and Lechner, M. (2002). 'Microeconometric evaluation of the active labour market policy in Switzerland', Economic Journal, vol. 112, pp. 854–93.

Greenberg, D.H., Michalopoulos, C. and Robins, P.K. (2003). 'A meta-analysis of government-sponsored training programs', *Industrial and Labor Relations Review*, vol. 57, pp. 31–53.

Hagglund, P. (2007). 'Are there pre-programme effects of Swedish active labour market policies? Evidence from three randomized experiments', Swedish Institute for Social Research Working Paper No. 2/2007, Stockholm: Stockholm University.

Heckman, J.J., Ichimura, H., Smith, J.A. and Todd, P. (1998). 'Characterizing selection bias using experimental data', *Econometrica*, vol. 66, pp. 1017–98.

Heckman, J.J., Ichimura, H. and Todd, P. (1998). 'Matching as an econometric evaluation estimator', *Review of Economic Studies*, vol. 65, pp. 261–94.

Heckman, J.J., Lalonde, R.J. and Smith, J.A. (1999). 'The economics and econometrics of active labour market programs', in (O. Ashenfelter and D. Card, eds), *Handbook of Labour Economics*, Volume 3A. pp. 1865–2095, Amsterdam and New York: Elsevier.

Heckman, J.J. and Robb, R. (1985). 'Alternative methods for evaluating the impact of interventions', in (J.J. Heckman and B. Singer, eds), *Longitudinal Analysis of Labour Market Data*, pp. 156–246, Cambridge: Cambridge University Press.

Hedges, L.V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*, New York: Academic Press.

Higgins, J.P.T. and Green, S. (editors) (2008). *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.0.1, available at http://www.cochrane-handbook.org.

Hotz, V.J., Imbens, G. and Klerman, J (2006). 'Evaluating the differential effects of alternative welfare-to-work training components: a re-analysis of the California GAIN program', *Journal of Labour Economics*, vol. 24, pp. 521–66.

Imbens, G. and Wooldridge, J.M. (2009). 'Recent developments in the econometrics of program evaluation', *Journal of Economic Literature*, vol. 47, pp. 5–86.

Jespersen, S.T., Munch, J.R. and Skipper, L. (2008). 'Costs and benefits of Danish active labour market programmes', *Labour Economics*, vol. 15, pp. 859–84.

Johnson, G.P. (1976). 'Evaluating the macroeconomic effects of public employment programs', in (O. Ashenfelter and J. Blum, eds), *Evaluating the Labor Market Effects of Social Programs*, Princeton, NJ: Princeton University Industrial Relations Section.

Kluve, J. (forthcoming). 'The effectiveness of European active labour market programs', *Labour Economics*.

Kluve, J. and Schmidt, C.M. (2002). 'Can training and employment subsidies combat European unemployment?', *Economic Policy*, vol. 35, pp. 409–48.

Lalonde, R.J. (1986). 'Evaluating the econometric evaluations of training programs with experimental data', *American Economic Review*, vol. 76, pp. 604–20.

Lechner, M. and Wunsch, C. (2006). 'Active labour market policy in East Germany: waiting for the economy to take off', IZA Working Paper No. 2363. Institute for the Study of Labour(IZA), Bonn.

Richardson, K. and van den Berg, G.J. (2002). 'The effect of vocational employment training on the individual transition rate from unemployment to work', IFAU Working Paper 2002:8. Institute for Labour Market Policy Evaluation, Uppsala.

Sianesi, B. (2004). 'An evaluation of the Swedish system of active labour market programs in the 1990s', *Review of Economics and Statistics*, vol. 86, pp. 133–55.