# WHAT WORKS? A META ANALYSIS OF RECENT ACTIVE LABOR MARKET PROGRAM EVALUATIONS

**David Card**
UC Berkeley

**Jochen Kluve**
Humboldt University, RWI Essen

**Andrea Weber**
Central European University, WU Vienna

**Abstract**
We summarize the estimates from over 200 recent studies of active labor market programs. We classify the estimates by type of program and participant group, and distinguish between three different post-program time horizons. Using regression models for the estimated program effect (for studies that model the probability of employment) and for the sign and significance of the estimated effect (for all the studies in our sample) we conclude that: (1) average impacts are close to zero in the short run, but become more positive 2–3 years after completion of the program; (2) the time profile of impacts varies by type of program, with larger average gains for programs that emphasize human capital accumulation; (3) there is systematic heterogeneity across participant groups, with larger impacts for females and participants who enter from long term unemployment; (4) active labor market programs are more likely to show positive impacts in a recession. (JEL: J00, J68)

## 1. Introduction

In the long period of recovery after the Great Recession there is renewed interest in the potential use of active labor market policies (ALMPs) to help ease a wide range of labor market problems, including youth unemployment and persistent joblessness among displaced adults (e.g., Martin 2014). Although training programs, employment subsidies, and similar policies have been in use for well over 50 years, credible evidence on their causal impacts has only become available in recent decades (see Lalonde 2003

*The editor in charge of this paper was M. Daniele Paserman.*

Acknowledgments: We are extremely grateful to the editor and five referees for helpful comments on an earlier draft, and to seminar participants at IRVAPP Trento, ILO Geneva, OECD Paris, European Commission Brussels, The World Bank Washington DC, University of Oslo, ISF Oslo, MAER-Net 2015 Prague Colloquium, IFAU Uppsala. We also thank Diana Beyer, Hannah Frings and Jonas Jessen for excellent research assistance. Financial support from the Fritz Thyssen Foundation and the Leibniz Association is gratefully acknowledged. Card is a Research Associate at NBER.

E-mail: card@econ.berkeley.edu (Card); jochen.kluve@hu-berlin.de (Kluve); WeberA@ceu.edu (Weber)

*Journal of the European Economic Association* 2018 16(3):894–931 DOI: 10.1093/jeea/jvx028
© The Authors 2017. Published by Oxford University Press on behalf of European Economic Association. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

for a brief history). Within a relatively short period of time the number of scientific evaluations has exploded, offering the potential to learn what types of programs work best, in what circumstances, and for whom.

In this paper we synthesize the recent ALMP evaluation literature, looking for systematic evidence on these issues.[1] We extend the sample used in our earlier analysis (Card, Kluve, and Weber 2010; hereafter CKW), doubling the number of studies (from 97 to 207) and increasing the number of separate program estimates from 343 to 857. Many of the latest ALMP studies measure impacts on the employment rate of participants, yielding over 350 estimates for this outcome that can be readily compared across studies.

This new sample of estimates allows us extend our earlier work in 4 main ways. First, we can more precisely characterize average program impacts by type of ALMP and post-program time horizon. Second, we are able to compare the relative efficacy of different types of ALMP's (e.g., training versus job search assistance) for different participant groups (e.g., youths versus older workers). Third, we provide new evidence on the variation in program effects at different points in the business cycle. Finally, we conduct a systematic analysis of potential publication biases in the recent ALMP literature.

We summarize the estimates from different studies in two complementary ways. Our main approach is to examine the estimated program effects on employment, ignoring the findings from studies that model other outcomes (such as the duration of time to an unsubsidized job). Our second approach—which can be applied to all the estimates in our sample, regardless of the outcome variable—is to classify "sign and significance" based on whether the estimated impact is significantly positive, statistically insignificant, or significantly negative. The narrower focus of the first approach is preferred in the meta-analysis literature (e.g., Hedges and Olkin 1985; Roberts and Stanley 2005; Stanley and Doucouliagos 2012), because the magnitude of the effect is not mechanically related to the number of observations used in the study, whereas statistical significance is (in principle) sample-size dependent. Fortunately, the two approaches yield similar conclusions when applied to the subset of studies for which employment effects are available, giving us confidence that our main findings are invariant to how we summarize the literature.

We reach four main substantive conclusions. First, consistent with the pattern documented in CKW, we find that ALMPs have relatively small average effects in the short run (less than a year after the end of the program), but larger average effects in the medium run (1–2 years post program) and longer run (2+ years). Across studies that model impacts on employment, the short run impacts are centered between 1 and

1. Previous reviews include Heckman, LaLonde, and Smith (1999), who summarize 75 microeconometric evaluations from the United States and other countries, Kluve (2010), who reviews close to 100 studies from Europe, and Filges et al. (2015), who analyze a narrower set of 39 studies. Greenberg, Michalopoulos and Robins (2003) review U.S. programs targeted to disadvantaged workers. Bergemann and van den Berg (2008) survey program effects by gender. Ibarrarán and Rosas (2009) review programs in Latin America supported by the Inter-American Development Bank. Related meta analyses focusing on labor market interventions in low and middle income countries include Cho and Honorati (2014) and Grimm and Paffhausen (2015).

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

3 percentage points (ppt.) The distribution of medium run effects is shifted to the right, centered around 3–5 ppt., whereas the longer run effects are centered between 5 and 12 ppt. As a benchmark, the gap in employment rates between U.S. men with only a high school education and those with a 2 or 3 year community college degree is 10 ppts., suggesting that a 5–10 ppt. longer run impact is economically meaningful.[2]

Second, the time profile of average impacts in the post-program period varies with the type of ALMP. Job search assistance programs that emphasize "work first" tend to have similar impacts in the short and long run, whereas training and private sector employment programs have larger average effects in the medium and longer runs. Public sector employment subsidies tend to have small or even negative average impacts at all horizons.

Third, we find that the average impacts of ALMPs vary across groups, with larger average effects for females and participants drawn from the pool of long term unemployed, and smaller average effects for older workers and youths. We also find suggestive evidence that certain programs work better for specific subgroups of participants. Job search assistance programs appear to be relatively more successful for disadvantaged participants, whereas training and private sector employment subsidies tend to have larger average effects for the long term unemployed. Finally, comparing the relative efficacy of ALMPs offered at different points in the business cycle, we find that programs in recessionary periods tend to have larger average impacts, particularly if the downturn is relatively short-lived.

On the methodological side, we find that the average program effects from randomized experiments are not very different from the average effects from nonexperimental designs. This is reassuring given longstanding concerns over the reliability of nonexperimental methods for evaluating job training and related programs (e.g., Ashenfelter 1987). We also find that there is substantial unobserved heterogeneity in the estimated program impacts in the literature. This heterogeneity is large relative to the variation attributable to sampling error, leading to relatively wide dispersion in the estimated impacts from designs with similar precision. In contrast to the patterns uncovered in meta-analyses of minimum wage effects (Doucouliagos and Stanley 2009) and the intertemporal substitution elasticity (Havránek 2015) this dispersion is also nearly symmetric. As a result, standard tests for publication bias, which look for asymmetry in the distribution of program estimates, are insignificant.

## 2. Sample Construction

### *2.1. Sampling Impact Evaluation Studies*

We extend the sample in CKW, using the same criteria to select in-scope studies and the same protocols to extract information about program features and impacts. The

2. In 2015, the average monthly employment rate of men over age 25 with a high school education in the United States was 63.5%; the average for men with an associate degree was 73.8% (United States Department of Labor 2016).

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

CKW sample was derived from responses to a 2007 survey of researchers affiliated with the Institute for the Study of Labor (IZA) and the National Bureau of Economic Research (NBER) asking about evaluation studies written after 1995.[3] To extend this sample we began by reviewing the research profiles and homepages of IZA research fellows with a declared interest in "program evaluation", looking for studies written since 2007. We also searched the NBER working paper database using the search strings "training", "active", "public sector employment", and "search assistance."

In a second step we used a Google Scholar search to identify all papers citing CKW or the earlier review by Kluve (2010). We also searched through the *International Initiative for Impact Evaluation*'s "Repository of Impact Evaluation Published Studies," the online project list of the *Abdul Latif Jameel Poverty Action Lab* (J-PAL), and the list of Latin American program evaluations reviewed by Ibarrarán and Rosas (2009).

After identifying an initial sample of studies, we reviewed the citations in all the papers to find any additional ALMP studies. We also identified four additional papers presented at a conference in early fall 2014. The search process lasted from April to October 2014 and yielded 154 new studies that were considered for inclusion in our ALMP impact evaluation data base.

### *2.2. Inclusion Criteria*

In order to generate a consistent data base across the two waves of data collection (2007 and 2014) we imposed the same restrictions adopted in CKW. First, the program(s) analyzed in the evaluation had to be one of following five types:

- classroom or on-the-job training
- job search assistance, monitoring, or sanctions for failing to search
- subsidized private sector employment
- subsidized public sector employment
- other programs combining two or more of the above types.[4]

Since our focus is on "active" labor market policies, we exclude studies of financial incentives, such as re-employment bonuses (summarized in Meyer 1995) or earnings subsidy programs (discussed in Blank, Card and Robins 2000). We also exclude open-ended entitlement programs like child care subsidies, and include only *individually targeted* employer subsidy programs, excluding tax incentives or other subsidies that are available for all newly hired or existing workers. Finally, we exclude studies that revise or update an older study in the CKW sample, or have substantial overlap with

3. The 1995 starting point was determined in part by the existence of several well-known summaries of the literature up to the mid-1990s, including Friedlander, Greenberg, and Robins (1997), Heckman et al. (1999), and Greenberg et al. (2003).

4. Most of these programs combine an element of job search with training or subsidized employment. We also include 7 estimates of the "threat of assignment" to a program in this category.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

an older study. Methodologically, we include only well-documented studies that use individual micro data and incorporate a counterfactual/control group design or some form of selection correction.

Imposing these criteria we retain 110 of the 154 studies identified in the search process.[5] We added these to the 97 studies from CKW, yielding a final sample of 207 impact evaluations. A complete list of these studies is contained in the Online Appendix B and our entire data base of program estimates is downloadable on the journal webpage.

We emphasize that the evaluations in our sample have many limitations. At best, these studies measure the partial equilibrium effects of ALMPs, comparing the mean outcomes in a treatment group to those of an untreated control or comparison group.[6] Even from this narrow perspective few studies present information on the costs of the program, and detailed cost–benefit calculations are very rare. Moreover, although we restrict attention to studies with a comparison group or selection correction design, we suspect that there may be some bias in the estimates from any particular study. We do not believe, however, that authors have a strong incentive to choose specifications that lead to positive program estimates, since many well-known studies in the literature report insignificant or even negative impacts for some programs or subgroups (e.g., Bloom et al. 1997). Thus, we do not have a strong presumption that the biases in the literature tend to be one-sided.

### *2.3. Extracting Impact Estimates and Information on Programs and Participants*

The next step was to extract information about the programs and participants analyzed in each study, and the corresponding program impact estimates.[7] Using the classification system developed in CKW, we gathered information on the type of ALMP, on the types of participants that are admitted to the program (long term unemployed, regular unemployment insurance recipients, or disadvantaged individuals[8]), the type of dependent variable used to measure the impact of the program, and the econometric methodology. We also gathered information on the (approximate) dates of operation of the program, the age and gender of participants in the program, the source of the data used in the evaluation (administrative records or a specialized survey), and the approximate duration of the program.

If a study reported separate impact estimates either by program type or by participant group, we identified the *program/participant subgroup* (PPS) and coded

5. The main reasons for exclusion were: overlap with other papers (i.e., estimating impacts for the same program); program out of scope; and no explicit counterfactual design.

6. The literature on the equilibrium effects of ALMP is scarce. For a notable exception, see Crépon et al. (2013).

7. As in CKW, we extracted the information from the studies ourselves, since we found that substantial knowledge of evaluation methodology and the ALMP literature is often needed to interpret the studies.

8. We classify the intake group as "disadvantaged" if participants are selected from low-income or low-labor market attachment individuals.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

the impact estimates separately. Overall, we have information on 526 separate PPSs from the 207 studies, with a minimum of 1 and a maximum of 10 PPSs in each study. We also identified up to three impact estimates for each PPS, corresponding to three different post-program time horizons: short-term (approximately one year after completion of the program); medium term (approximately 2 years after); and longer-term (approximately 3 years after). In total, we have 857 separate program estimates for the 526 program/participant subgroups, with between one and three estimates of the effect of the program at different time horizons.[9]

We use two complementary approaches to quantify the estimated program impacts. First, we classify the estimates as significantly positive, insignificantly different from zero, or significantly negative (at the 5% level). This measure of effectiveness is available for every estimate in our data base. For the subset of studies that measure effects on the probability of employment, we also extract an estimate of the *program effect* on the employment rate of participants.[10]

The final step in our data assembly procedure was to add information on labor market conditions at the time of operation of the program. Specifically, we gathered information on GDP growth rates and unemployment rates from the OECD, the World Bank, and the ILO. For our main analysis we focus on how program effectiveness is related to the average growth rate and the average unemployment rate during the period the program group participated in the ALMP, though we also look at the effect of conditions in the post-program period.

## 3. Descriptive Overview

### *3.1. Program Types, Participant Characteristics, Evaluation Design*

Table 1 presents an overview of the program estimates in our final sample. As noted, we have a total of 857 different impact estimates for 526 different PPSs (program-type/participant subgroup combinations) extracted from 207 separate studies. To deal with potential correlations between the program estimates from a given study—arising for example from idiosyncratic features of the evaluation methodology—we calculate standard errors *clustering by study*.

Column (1) presents the characteristics of our overall sample, whereas columns (2)–(6) summarize the estimates from five country groups: the Germanic countries (Austria, Germany, and Switzerland), which account for about one quarter of all studies; the Nordic countries (Denmark, Finland, Norway, and Sweden), which account for another quarter of studies; the Anglo countries (Australia, Canada, New Zealand,

9. For a specific PPS and time horizon we try to identify and code the main estimate in the study. We do not include multiple estimates for the same PPS and time horizon.

10. We also extract the average employment rate of the comparison group, and for some analysis we model the program effect divided by either the comparison group employment rate or the standard deviation of the comparison group employment rate.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

TABLE 1. Description of sample of program estimates.

| | Full sample (1) | Country Group: Austria, Germany, Switzerland (2) | Nordic Countries (3) | U.S., U.K, Aust., N.Z., Canada (4) | Non-OECD (5) | Latin Amer. and Caribbean (6) |
|---|---|---|---|---|---|---|
| Number of estimates | 857 | 290 | 212 | 87 | 132 | 72 |
| Number of PPS's | 526 | 163 | 127 | 45 | 86 | 54 |
| Number of studies | 207 | 52 | 48 | 24 | 33 | 19 |
| *Type of program (%)* | | | | | | |
| Training | 49 | 62 | 17 | 45 | 79 | 97 |
| Job search assistance | 15 | 8 | 26 | 22 | 2 | 0 |
| Private subsidy | 14 | 17 | 15 | 5 | 11 | 3 |
| Public employment | 9 | 9 | 10 | 3 | 6 | 0 |
| Other | 14 | 5 | 32 | 25 | 2 | 0 |
| *Age of program group (%)* | | | | | | |
| Mixed | 59 | 54 | 61 | 72 | 40 | 25 |
| Youth (<25 years) | 21 | 12 | 20 | 15 | 53 | 69 |
| Older (≥25 years) | 20 | 33 | 19 | 13 | 8 | 6 |
| *Gender of program group (%)* | | | | | | |
| Mixed | 54 | 53 | 67 | 43 | 43 | 11 |
| Males only | 22 | 24 | 18 | 25 | 23 | 44 |
| Females only | 23 | 23 | 16 | 32 | 31 | 44 |
| *Type of program participants (%)* | | | | | | |
| Registered unemployed | 65 | 86 | 67 | 33 | 24 | 0 |
| Long-term unemployed | 12 | 8 | 10 | 25 | 7 | 0 |
| Disadvantaged | 23 | 6 | 23 | 41 | 69 | 100 |
| *Outcome of interest (%)* | | | | | | |
| Employment status | 57 | 83 | 31 | 26 | 63 | 54 |
| Earnings | 21 | 8 | 25 | 47 | 36 | 43 |
| Hazard to new job | 12 | 7 | 25 | 3 | 0 | 0 |
| Other hazard | 6 | 0 | 16 | 2 | 0 | 3 |
| Unemployment status | 4 | 2 | 4 | 21 | 1 | 0 |
| *Effect measured at (%)* | | | | | | |
| Short term | 48 | 42 | 54 | 37 | 47 | 57 |
| Medium term | 35 | 34 | 31 | 40 | 45 | 42 |
| Long term | 16 | 23 | 16 | 23 | 8 | 1 |
| Experimental design (%) | 19 | 0 | 39 | 31 | 28 | 26 |

Notes: See text for description of sample. Study refers to an article or unpublished paper. PPS refers to a program/participant subgroup (e.g., a job search assistance program for mixed gender youths). Estimate refers to an estimate of the effect of the program on the participant subgroup at either a short-term (<1 year after completion of the program), medium term (1–2 years post completion) or long term (2+ years post completion) time horizon. Job search assistance programs include sanction programs. "Other" programs include those that combine elements of the four distinct types.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

United Kingdom, and United States), which account for just over 10% of studies; and two nonmutually exclusive groups of lower/middle income countries—"non-OECD" countries (10% of studies), and Latin American and Caribbean (LAC) countries (10% of studies). Online Appendix Figure A.1 shows the numbers of estimates by country. The largest source countries are Germany (253 estimates), Denmark (115 estimates), Sweden (66 estimates), the United States (57 estimates), and France (42 estimates).

The second panel of Table 1 shows the distribution of program types in our sample. Training programs (including classroom and on-the-job training) account for about one half of the program estimates, with bigger shares in the non-OECD and LAC countries. Public sector employment programs, by comparison, are relatively rare among recent evaluations, whereas job search assistance (JSA) programs, private employment subsidies, and other/combined programs each represent about 15% of the estimates.[11]

The next three panels of the table show the characteristics of the program participants, classified by age group, gender, and "type" of participant. About one-half of the estimates are for mixed age and mixed gender groups, but we also have relatively large subsets of estimates that are specific to either younger or older workers, or females or males. Sixty-five percent of the program estimates (and nearly all the estimates from the Germanic countries) are for participants who enter from the unemployment insurance (UI) system. Typically these participants are assigned to a program and required to attend as a condition for continuing benefit eligibility.[12] The remaining 35% of estimates are split between programs that serve the long term unemployed (LTU) and those that serve disadvantaged participant groups. In many cases, these groups are recruited by program operators and enroll voluntarily. Such voluntary programs are more common in the Anglo Saxon countries and in less developed countries that lack a formal UI system.[13]

Next we show the outcome variables used to measure the program impact and the time horizons of the estimate. The most common outcome—particularly in the Germanic and non-OECD countries—is the probability of employment, whereas the level of earnings is the most common metric in the Anglo Saxon countries. About one sixth of the program estimates—but 40% of those from Nordic countries—measure the exit rate from the benefit system, typically focusing on the rate of exit to a new (unsubsidized) job. Finally, a small subset of estimates—mostly from Anglo Saxon countries—focus on the probability of unemployment. About one half of the estimates are for a short term horizon (<1 year) after program completion, 35% for a medium term (1–2 years), and 16% for a longer term (more than 2 year after).

11. The JSA category includes a small number of evaluations (with a total of 8 program estimates) for programs that monitor search activity and threaten sanctions for low search effort. We combine these with JSA programs because both types of programs have similar incentive effects on participants' search activity.

12. This type of program requirement is widespread in Europe—see Sianesi (2004) for a discussion.

13. The U.S. job training programs analyzed in the seminal papers of Ashenfelter (1978), Ashenfelter and Card (1985), Lalonde (1986), Heckman et al. (1998) are all of this type.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

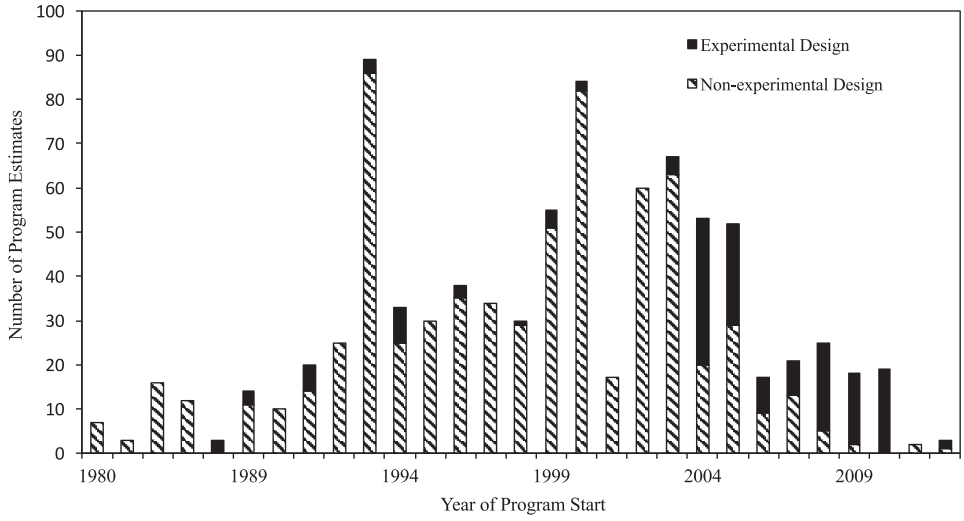


FIGURE 1. Number of program estimates, by year of program start.

The last row of the table shows the fraction of program estimates that are based on an experimental design. In most of our country groups about 30% of estimates come from randomized controlled trials (RCTs) that have been explicitly designed to measure the effectiveness of the ALMP of interest. An important exception is the Germanic countries, where no experimentally based estimates are yet available.

The distribution of program estimates over time (defining time by the earliest intake year of the program) is shown in Figure 1, with separate counts for experimental and nonexperimental estimates. Our sample includes programs from as far back as 1980, though the majority of estimates are from the 1990s and early 2000s, reflecting our focus on studies written since 1995. There is clear evidence of a trend toward increasing use of experimental designs: among the 210 estimates from 2004 and later, 61% are from randomized designs.

### *3.2. Measures of Program Impact—Overview*

Table 2 gives an overview of our two main measures of program impact, contrasting results for the short term, medium term, and long term. Column (1) summarizes the sign and significance of all the available program estimates. Among the 415 short term estimates, 40% are significantly positive, 42% are insignificant, and 18% are significantly negative. The pattern of results is more positive in the medium and longer terms, with a majority of estimates (52%) being significantly positive in the medium term, and 61% being significantly positive in the longer term.

Column (2) shows the distribution of sign and significance for the subset of studies that use post-program employment rates to evaluate the ALMP program. These 111 studies account for 490 program estimates (57% of our total sample). The short

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

TABLE 2. Summary of program estimates by availability of estimate of program effect on probability of employment.

| | Sign and significance of program effects | | | Estimated program effect on prob. of emp. (col. (3) subsample) | |
|---|---|---|---|---|---|
| | Full sample (1) | Subsample with outcome = prob. of emp. (2) | Subsample with estimate of effect on prob. of emp. (3) | Mean effect (std. error)[a] (4) | Median effect (std. error)[a] (5) |
| Number of estimates | 857 | 490 | 352 | 352 | |
| Number of PPS's | 526 | 274 | 200 | 200 | |
| Number of studies | 207 | 111 | 83 | 83 | |
| *Short term estimates* | | | | | |
| All ST estimates—number [pct. of total estimates] | 415 [48] | 205 [42] | 141 [40] | 1.6 (0.8) | 1.0 (1.0) |
| Significant positive ST estimate—pct. of ST sample | 40 | 31 | 33 | 8.8 (1.3) | 6.0 (1.3) |
| Insignificant ST estimate—pct. of ST sample | 42 | 47 | 44 | 0.5 (0.4) | 0.0 (0.6) |
| Significant negative ST estimate—pct. of ST sample | 18 | 22 | 23 | −6.4 (0.8) | −5.0 (0.7) |
| *Medium term (MT) estimates* | | | | | |
| All MT estimates—number [pct. of total estimates] | 301 [35] | 194 [40] | 143 [41] | 5.4 (1.2) | 3.0 (0.7) |
| Significant positive MT estimate—pct. of MT sample | 52 | 50 | 47 | 11.3 (1.9) | 8.5 (1.1) |
| Insignificant MT estimate—pct. of MT sample | 40 | 41 | 43 | 1.3 (0.3) | 1.0 (0.4) |
| Significant negative MT estimate—pct. of MT sample | 8 | 9 | 10 | −5.0 (1.2) | −4.9 (2.2) |
| *Long term (LT) estimates* | | | | | |
| All LT estimates—number [pct. of total estimates] | 141 [16] | 91 [19] | 68 [19] | 8.7(2.2) | 4.9 (1.4) |
| Significant positive LT estimate—pct. of sample | 61 | 65 | 65 | 13.0 (2.7) | 9.0 (2.2) |
| Insignificant LT estimate—pct. of sample | 35 | 32 | 32 | 1.3 (0.6) | 1.1 (0.7) |
| Significant negative LT estimate—pct. of sample | 4 | 3 | 3 | −4.2 (0.5) | – |

Notes: See note to Table 1. Short term program estimates are for the period up to 1 year after the completion of the program. Medium term estimates are for the period from 1 to 2 years after completion of the program. Long term estimates are for the period 2 or more years after completion of the program. Effect sizes are only available for studies that model the probability of employment as the outcome of interest, and provide information on mean employment rate of comparison group.

a. Standard errors are clustered by study.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

term program estimates from this subset of studies are somewhat less positive than in the overall sample. In the medium and longer terms, however, the discrepancy disappears. As discussed below, these patterns are not explained by differences in the types of ALMP programs analyzed in different studies, or by differences in participant characteristics. Instead, they reflect a tendency for studies based on models of the time to unemployment exit (which are included in column (1) but excluded in column (2)) to exhibit more positive short term impacts than studies based on employment.

Column (3) of Table 2 shows the distributions of sign and significance associated with the estimated employment effects where we can extract both an actual program effect (typically the coefficient from a linear probability model) *and* the employment rate of the comparison group.[14] The distributions are very similar to those in column (2), suggesting that there is no systematic bias associated with the availability of an impact effect and the comparison group employment rate.

Finally, columns (4) and (5) report the mean and median of the distributions of estimated program effects for the subsample in column (3). The short run program effects are centered just above zero, with a mean and median of 1.6 and 1.0 ppt., respectively. In the medium term the distribution shifts right but also becomes slightly more asymmetric, with a mean and median of 5.4 and 3.0 ppt., respectively. In the long term there is a further shift right, particularly in the upper half of the distribution, with a mean and median of 8.7 and 4.9 ppt., respectively.

Positive skew in the distribution of estimated effects is often interpreted in the meta-analysis literature as evidence of "publication bias", particularly if the positive effects are imprecisely estimated (see, e.g., Stanley and Doucouliagos 2012). Some insight into this issue is offered by the "forest plots" in Figure 2(a)–(c), which show the cumulative distributions of program estimates at each time horizon, along with bands representing the standard errors of the estimates.[15]

Inspection of these graphs confirms that the overall distribution of program effects shifts to the right as the time horizon is extended. At all three horizons there is also some positive skew in the distribution of estimated effects. Interestingly, however, the confidence intervals do not appear to be systematically wider for estimates in the upper or lower tails of the distribution. Instead, there are a handful of positive outliers in the short and medium term distributions that push the unweighted mean above the median and the precision-weighted mean.

Returning to Table 2, columns (4) and (5) also show the mean and median program effects for estimates that are classified as significantly positive, insignificant, or significantly negative. As would be expected if differences in sign and significance are mainly driven by differences in the magnitude of the program estimates—rather than by

14. In many cases a study reports the impact on the employment rate of the program group but does not report the employment rate of the comparison group. As discussed below, we need the latter number to construct effect sizes or proportional impacts on the employment rate.

15. The distributions are limited to estimates for which we also have an estimate of the associated standard error. Information on the standard errors of program estimates was not extracted in CKW. Thus, the estimates are from the latest studies collected in our second round.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

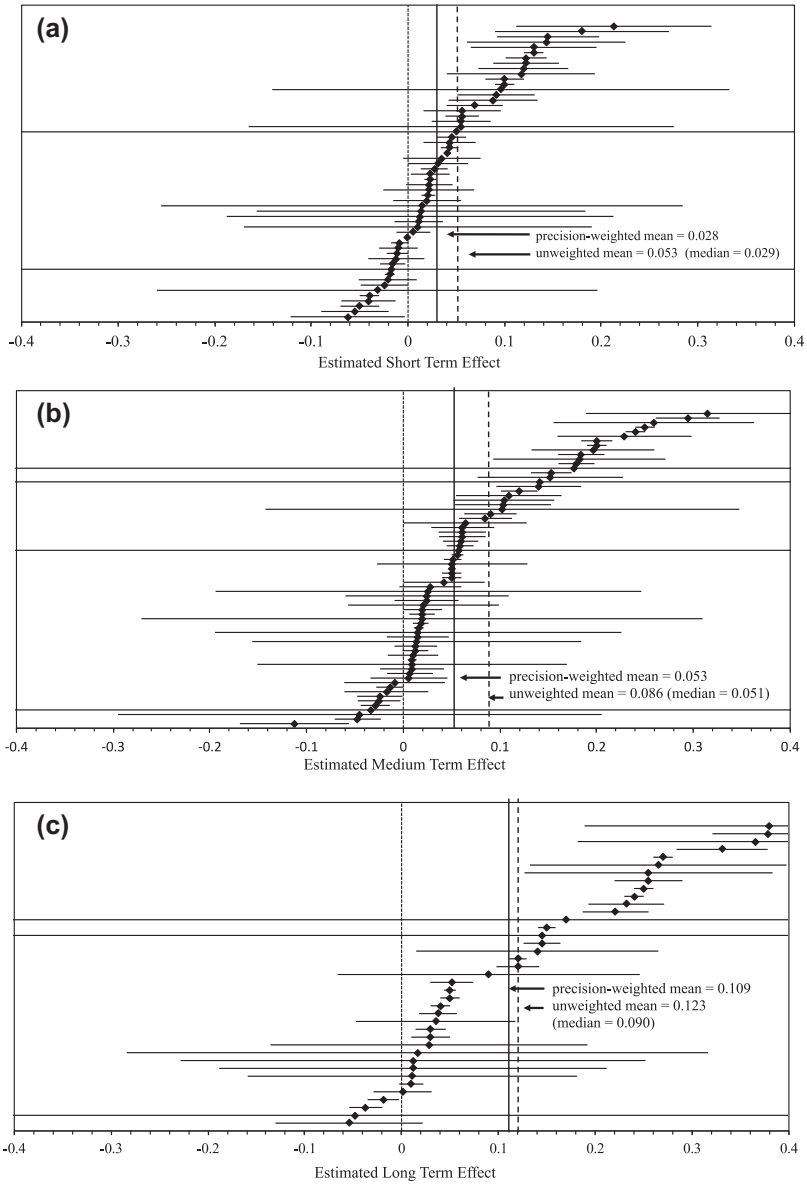


FIGURE 2. (a) Short term effects and confidence intervals. Diamonds represent estimated short term treatment effects on probability of employment for a program/participant subgroup (PPS). Horizontal lines represent 95% confidence intervals. Graph shows 56 estimates—2 large positive estimates are not shown for clarity. (b) Medium term effects and confidence intervals. Diamonds represent estimated medium term treatment effects on probability of employment for a program/participant subgroup (PPS). Horizontal lines represent 95% confidence intervals. Graph shows 69 estimates—3 large positive estimates are not shown for clarity. (c) Long term effects and confidence intervals. Diamonds represent estimated long term treatment effects on probability of employment for a program/participant subgroup (PPS). Horizontal lines represent 95% confidence intervals. Graph shows 39 estimates.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

differences in the standard errors of the estimates—the mean and median are large and positive for significant positive estimates, large and negative for significant negative estimates, and close to zero for insignificant estimates. This pattern is illustrated in Online Appendix Figures A.2(a)–(c), where we plot the histograms of estimated effects at each time horizon, separating the estimates by category of sign and significance. At all three time horizons, the subgroups of estimates appear to be drawn from distributions that are centered on different midpoints. This separation suggests that the sign and significance of an estimate can serve as noisy indicator of the underlying effect.

### *3.3. Variation in Average Program Impacts*

Tables 3(a) and 3(b) provide a first look at the question of how average ALMP impacts vary across different types of programs and different participant groups. For each subset of estimates we show the mean program effects at each time horizon and the corresponding fraction of program estimates that is significantly positive.

Focusing first on comparisons across program types (Table 3(a)), notice that training and private sector employment programs tend to have small average effects in the short run, coupled with more positive average impacts in the medium and longer runs. In contrast, JSA programs and ALMPs in the "other" category have more stable impacts. These profiles are consistent with the nature of the two broad groups of programs. Participants in training and private subsidy programs often suspend their normal job search efforts and devote their time to program activities—a so-called "lock-in" effect that typically leads to worse outcomes in the immediate post-program period (see, e.g., Ham and Lalonde 1996).[16] Assuming that investments made during the program period are valuable, however, the outcomes of participants will gradually catch up with those of the comparison group.[17] By comparison, JSA programs and other programs that include monitoring of search are designed to push participants into the labor market quickly, with little or no investment component. In the absence of large returns to recent job experience, it is unlikely that these programs can have large long run effects.[18]

Another clear finding in Table 3(a) is the relatively poor performance of public sector employment programs—a result that has been found in other previous analyses (e.g., Heckman et al. 1999, and CKW).

Online Appendix Figure A.3(a) shows how the relative share of different types of ALMPs have changed over the 30 year period covered by our sample. The

16. In cases where the program group is drawn from the regular UI system, participants in training and subsidized employment opportunities are often exempt from search requirements that are imposed on nonparticipants—see, for example, Biewen et al. (2014) for a discussion in the German context.

17. As noted by Mincer (1974) a similar crossover pattern is observed in the comparison of earnings profiles of high school graduates and college graduates.

18. Evidence on the value of labor market experience for lower skilled workers (Gladden and Taber 2000; Card and Hyslop 2005) suggests that the returns are modest and unlikely to exceed 2% or 3% per year of work.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

TABLE 3(a). Comparison of impact estimates by program type and participant group.

| | Number Est's. (1) | Median sample size (2) | Percent RCT's (3) | Mean program effect on prob. emp. (×100) Short term (4) | Medium term (5) | Longer term (6) | Pct. of estimates with sig. positive impact Short term (7) | Medium term (8) | Longer term (9) |
|---|---|---|---|---|---|---|---|---|---|
| All | 857 | 10,709 | 19.4 | 1.6 | 5.4 | 8.7 | 40 | 52 | 61 |
| | | | | (141) | (143) | (68) | (415) | (301) | (141) |
| *By program type* | | | | | | | | | |
| Training | 418 | 7700 | 12.9 | 2.0 | 6.6 | 6.7 | 35 | 54 | 67 |
| | | | | (90) | (92) | (35) | (201) | (163) | (54) |
| Job search assist. | 129 | 4648 | 51.2 | 1.2 | 2.0 | 1.1 | 53 | 63 | 43 |
| | | | | (16) | (13) | (7) | (68) | (40) | (21) |
| Private subsidy | 118 | 10,000 | 8.5 | 1.1 | 6.2 | 21.1 | 37 | 65 | 88 |
| | | | | (13) | (17) | (16) | (49) | (37) | (32) |
| Public sector emp. | 76 | 17,084 | 0.0 | 3.6 | −1.1 | 0.8 | 32 | 25 | 27 |
| | | | | (14) | (12) | (6) | (41) | (24) | (11) |
| Other | 116 | 17,391 | 31.0 | 7.2 | 5.8 | 2.0 | 52 | 38 | 43 |
| | | | | (8) | (9) | (4) | (56) | (37) | (23) |
| *By intake group* | | | | | | | | | |
| UI recipients | 554 | 11,000 | 17.1 | −0.1 | 4.3 | 8.5 | 34 | 47 | 59 |
| | | | | (93) | (101) | (50) | (258) | (193) | (103) |
| Long term unem. | 106 | 8900 | 16.0 | 5.8 | 13.0 | 12.7 | 50 | 65 | 63 |
| | | | | (17) | (16) | (10) | (50) | (40) | (16) |
| Disadvantaged | 197 | 7027 | 27.4 | 4.2 | 5.3 | 5.0 | 50 | 59 | 68 |
| | | | | (31) | (26) | (8) | (107) | (68) | (22) |

Notes: See Tables 1 and 2. Number of program estimates associated with each table entry is reported in parentheses. Program effects (columns (4)–(6)) are only available for studies that model the probability of employment as the outcome of interest.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

TABLE 3(b). Additional comparisons of impact estimates by participant groups and design.

| | Number Est's. (1) | Median sample size (2) | Percent RCT's (3) | Mean program effect on prob. emp. (×100) Short term (4) | Medium term (5) | Longer term (6) | Pct. of estimates with sig. positive impact Short term (7) | Medium term (8) | Longer term (9) |
|---|---|---|---|---|---|---|---|---|---|
| All | 857 | 10,709 | 19.4 | 1.6 | 5.4 | 8.7 | 40 | 52 | 61 |
| | | | | (141) | (143) | (68) | (415) | (301) | (141) |
| *By age* | | | | | | | | | |
| Mixed age | 505 | 10,000 | 16.6 | 1.7 | 6.7 | 10.5 | 47 | 57 | 65 |
| | | | | (71) | (84) | (51) | (238) | (178) | (89) |
| Youth (<25) | 180 | 3000 | 33.3 | 2.9 | 2.7 | 0.2 | 32 | 41 | 67 |
| | | | | (34) | (29) | (5) | (92) | (64) | (24) |
| Non-youth | 172 | 25,850 | 12.8 | 0.1 | 4.5 | 4.6 | 31 | 51 | 43 |
| | | | | (36) | (30) | (12) | (85) | (59) | (28) |
| *By gender* | | | | | | | | | |
| Mixed gender | 466 | 11,000 | 19.7 | 1.6 | 4.4 | 5.6 | 39 | 52 | 59 |
| | | | | (89) | (85) | (45) | (224) | (155) | (87) |
| Males only | 191 | 10,000 | 15.2 | 1.4 | 6.1 | 13.1 | 41 | 50 | 58 |
| | | | | (24) | (28) | (9) | (95) | (72) | (24) |
| Females only | 200 | 8345 | 22.5 | 4.1 | 7.8 | 15.9 | 41 | 55 | 70 |
| | | | | (28) | (30) | (14) | (96) | (74) | (30) |
| *By evaluation design* | | | | | | | | | |
| Experimental | 166 | 1471 | 100.0 | 4.4 | 2.5 | 0.5 | 40 | 41 | 37 |
| | | | | (28) | (25) | (15) | (78) | (58) | (30) |
| Nonexperimental | 691 | 16,000 | 0.0 | 0.9 | 6.0 | 11.0 | 40 | 55 | 68 |
| | | | | (113) | (118) | (53) | (337) | (243) | (111) |

Notes: see Tables 1 and 2. Number of program estimates associated with each table entry is reported in parentheses. Program effects (columns (4)–(6)) are only available for studies that model the probability of employment as the outcome of interest.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

shares of training and JSA programs is relatively stable, whereas the share of public sector employment programs has fallen sharply, perhaps reflecting the more negative evaluation results that these programs have often received.

Online Appendix Figure A.3(b) shows the variation over time in our two measures of program impact. Overall, the sign and significance classifications of short term, medium term, and long term estimates are quite stable over time, with little indication that more recent programs are more or less likely to show significant positive results. There is more variability in the mean impacts on the probability of employment, with some evidence of an upward trend, particularly for the short and medium term impacts.

The middle rows of Table 3(b) compare the distributions of program effects by participant age group and gender. The results for PPSs that include all age groups are quite similar to the results for the overall sample, whereas the results for youth participants show a mixed pattern, with relatively small average program effects on employment at all time horizons (columns (4)–(6)), but more evidence of positive long-run impacts based on sign and significance (columns (7)–(9)). The differences across gender groups are more systematic and indicate that average estimated program effects are slightly larger at all time horizons for females (columns (4)–(6)) and have a higher probability of being significantly positive (columns (7)–(9)).

Finally, the bottom rows of Table 3(b) contracts results from evaluations based on randomized designs and nonexperimental designs. The comparisons of mean effects suggest that experimentally based estimates tend to be larger in the short run and decline over time, whereas nonexperimentally based estimates tend to become larger (more positive) over time. We caution that these simple "one way" contrasts must be interpreted carefully, however, because there are multiple sources of potential heterogeneity in the program impacts. For example, many of the experimental evaluations focus on JSA programs, whereas many of the nonexperimental evaluations focus on training programs. The meta-analysis models in Section 4 directly address this issue using a multivariate regression approach.

### *3.4. Profile of Post-Program Impacts*

Simple comparisons across the impact estimates in our sample suggest that ALMPs have more positive average effects in the medium and longer terms. To verify that this is actually true for a *given* program and participant subgroup—and is not simply an artifact of heterogeneity across studies—we examine the within-PPS evolution of impact estimates in Table 4.

Columns (1)–(3) show the changes in estimated program effects on the probability of employment for the subset of studies for which we observe both short and medium term estimates, medium and long term estimates, and short and long term estimates, respectively. Consistent with the simple cross-sectional comparisons, the within-PPS effects tend to increase as the time horizon is extended from the short run to the medium run, or from the short run to the long run. The average change between the medium and longer runs is essentially zero.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

TABLE 4. Changes in within-program effects over different time horizons.

| | Change in program effect on prob. of emp. | | | Change in sign/significance | | |
|---|---|---|---|---|---|---|
| | Short term to medium term (1) | Short term to long term (2) | Medium term to long term (3) | Short term to medium term (4) | Short term to long term (5) | Medium term to long term (6) |
| All | 0.021 | 0.024 | −0.006 | 0.231 | 0.250 | 0.020 |
| | (0.008) | (0.015) | (0.004) | (0.055) | (0.103) | (0.052) |
| *Number studies* | *105* | *43* | *47* | *225* | *100* | *102* |
| *By program type* | | | | | | |
| Training | 0.032 | 0.044 | −0.004 | 0.314 | 0.439 | 0.048 |
| | (0.008) | (0.018) | (0.005) | (0.072) | (0.085) | (0.049) |
| *Number studies* | *70* | *28* | *28* | *121* | *41* | *42* |
| Job search assist. | 0.003 | −0.001 | 0.000 | 0.265 | 0.143 | −0.111 |
| | (0.008) | (0.001) | (0.002) | (0.095) | (0.167) | (0.144) |
| *Number studies* | *10* | *7* | *7* | *34* | *21* | *18* |
| Private subsidy | −0.020 | −0.004 | −0.012 | 0.083 | 0.167 | −0.062 |
| | (0.049) | (0.078) | (0.013) | (0.150) | (0.267) | (0.068) |
| *Number studies* | *9* | *2* | *6* | *24* | *12* | *16* |
| Public sector emp. | 0.016 | −0.049 | −0.019 | 0.158 | −0.143 | −0.143 |
| | (0.014) | (0.049) | (0.019) | (0.170) | (0.494) | (0.285) |
| *Number studies* | *10* | *2* | *2* | *19* | *7* | *7* |
| Sanction/threat | 0.004 | −0.021 | −0.014 | 0.000 | 0.158 | 0.211 |
| | (0.012) | (0.008) | (0.006) | (0.108) | (0.182) | (0.092) |
| *Number studies* | *6* | *4* | *4* | *27* | *19* | *19* |

Notes: Change in estimated program effect in column (1) represents the difference between the estimated medium term and short term effects on the probability of employment for a given program and participant subgroup (PPS). Changes in columns (2) and (3) are defined analogously. Change in sign/significance in column (4) is defined as +1 if the short term estimate is significantly negative and the medium term estimate is insignificant, or if the short term estimate is insignificant and the medium term estimate is significantly positive; 0 if the sign and significance of the short term and medium term estimates is the same; and −1 if the short term estimate is significantly positive and the medium term estimate is insignificant, or if the short term estimate is insignificant and the medium term estimate is significantly negative. Changes in columns (5) and (6) are defined analogously. Standard deviations (clustered by study number) in parentheses.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

Comparing across program types it is clear that the pattern of rising impacts is driven by training programs, which show a relatively large gain in estimated program effects from the short term to the medium term. The patterns for the other types of programs suggest relatively constant or declining average program effects over the post-program time horizon. In particular, in contrast to the patterns in Table 3(a), there is no indication of a rise in impacts for private employment subsidy programs over time, suggesting that the gains in Table 3(a) may be driven by heterogeneity between studies. We return to this point below.

In columns (4)–(6) we examine the within-study changes in sign and significance for a broader set of studies. Here, we assign a value of +1 to PPS estimates that change from insignificant to significantly positive or from significantly negative to insignificant; −1 to estimates that change from significantly positive to insignificant or from insignificant to significantly negative; and 0 to estimates that have the same classification over time. This simple summary points to similar conclusions as the changes in estimated program effects, though JSA programs show more evidence of a rise in impacts from the short-run to the medium run in column (4) than the comparison of estimated effects on the probability of employment in column (1).

Online Appendix Tables A.1(a) and A.1(b) present full crosstabulations of sign/significance at the various post-program time horizons. As suggested by the simple classification system used in Table 4, most program estimates either remain in the same category, or become more positive over time.

## 4. Meta Analytic Models of Program Impacts

### *4.1. Conceptual Framework*

Consider an ALMP evaluation that models an outcome $y$ observed for members of both a participant group and a comparison group. Let $b$ represent the estimated impact of the program on the outcomes of the participants from a given evaluation design, and let $\beta$ represent the probability limit of $b$ (i.e., the estimate that would be obtained if the sample size for the evaluation were infinite). Under standard conditions the estimate $b$ will be approximately normally distributed with mean $\beta$ and some level of precision $P$ that depends on both the sample size for the evaluation and the design features of the study.[19] Therefore we can write:

$$b = \beta + P^{-1/2}z, \tag{1}$$

19. For example, in an experiment with 50% of the sample in the treatment group and no added covariates, $P = N/[2\sigma^2\ (1+\ \delta^2\ )]$, where $N$ is the sample size, $\sigma$ is the standard deviation of the outcome $y$ in the control group, and $\delta\sigma$ is the standard deviation of the outcome for the program group. In more complex designs such as difference in differences or instrumental variables the precision will be smaller.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

where $z$ is a realization from a distribution that will be close to $N(0, 1)$ if the sample size is large enough. The term $P^{-1/2}z$ has the interpretation of the realized sampling error that is incorporated in $b$.

Assume that the limiting program effect associated with a given study ($\beta$) can be decomposed as

$$\beta = X\alpha + \varepsilon, \tag{2}$$

where $\alpha$ is a vector of coefficients and $X$ captures the observed sources of heterogeneity in $\beta$, arising for example from differences in the type of program or the gender or age of the program participants. The term $\varepsilon$ represents fundamental heterogeneity in the limiting program effect arising from the particular way a program was implemented, or specific features of the program or its participants, or the nature of the labor market environment.

Equations (1) and (2) lead to a model for the observed program estimates of the form:

$$b = X\alpha + u, \tag{3}$$

where the error $u = \varepsilon + P^{-1/2}\, z$ includes *both* the sampling error in the estimate $b$ and the unobserved determinants of the limiting program effect for a given study. We use simple regression models based on equation (3) to analyze the program effects on the probability of employment that are available in our sample. We interpret these models as providing descriptive summaries of the variation in average program effects with differences in the observed characteristics of a given program and participant group *in our sample*. Recognizing the structure of the error component in (3) we prefer OLS estimation, which weights each estimated program effect equally, rather than precision-weighed estimation, which would be efficient under the assumption that $\varepsilon = 0$.[20] As we show below, in contrast to "classical" meta-analysis settings where each estimate is based on a clinical trial of the same drug, the variation in $\varepsilon$ appears to be particularly large for ALMP's, reflecting the wide range of factors that can potentially cause a program to be more or less successful.

For our full sample of program estimates we use (unweighted) ordered probit (OP) models for the 3-way classification of sign and significance of each estimate. Note that the $t$-statistic associated with the estimated impact $b$ is just the ratio of the estimate to the square root of its estimated sampling variance (which is the inverse of its estimated precision). Using equation (3), we can therefore write

$$\begin{aligned} t &= P^{1/2}b \\ &= P^{1/2}X\alpha + z + P^{1/2}\varepsilon. \end{aligned}$$

If the precision $P$ of the estimated program effects is constant across studies and there are no unobserved determinants of the limiting program effect (i.e., $\varepsilon = 0$) the

20. See Solon, Haider, and Wooldridge (2015) for a discussion of weighting.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

$t$-statistic will be normally distributed with mean $X\alpha'$ where $\alpha' = P^{1/2}\,\alpha$. In this case the coefficients from an OP model for whether the $t$ statistic is less than $-2$, between $-2$ and 2, or greater than 2 (i.e., the sign and significance of the estimated program effects) will be *strictly proportional* to the coefficients obtained from a regression model of the corresponding estimated program effects.

In our sample the estimated precision of the program estimates varies widely across studies, and there is clearly unobserved heterogeneity in the impacts.[21] Surprisingly, however, for studies that examine the probability of employment as an outcome the estimated coefficients from OLS models based on equation (3) and OP models for sign/significance are very nearly proportional, suggesting that the same observable factors that tend to raise the estimated program effects also tend to lead to more positive $t$ statistics. Our interpretation of this pattern is that the sampling error component of the program estimates is small relative to the variation due to observed and unobserved heterogeneity, so the $t$-statistic varies across studies in proportion to the relative magnitude of the estimated program effect. We therefore use the OP models to summarize the broader set of program estimates.

### *4.2. Basic Models for Program Effect and for Sign and Significance*

Table 5 presents the estimates from a series of regression models for our sample of estimated program effects on the probability of employment. We pool the effects for different post-program horizons and include dummies indicating whether the estimate is for the medium or long term (with short term estimates in the omitted group). The basic model in column (1) includes only these controls and a set of dummies for the type of program (with training programs in the omitted category). Consistent with the simple comparisons in Table 3(a), we find that the program estimates are larger in the medium and long run, and that public sector employment programs are associated with smaller program effects.

The model in column (2) introduces additional controls for the type of participant and study characteristics, which are reported in Table 7 and discussed below. These controls slightly attenuate the growth in program effects over longer post-program horizons and also reduce the magnitude of the JSA program effect from $-3.2$ ppts. (and significant) to $-0.1$ ppts (and insignificant).

Columns (3)–(5) introduce a parallel set of models that allow the time profiles of post-program impacts to vary with the type of program. In these specifications the "main effects" for each program type show the short term impacts relative to training programs (the omitted type), whereas the interactions of program type with medium term and long term dummies show how the impacts evolve *relative to the profile for training programs* (which are summarized by the main effects in the first two rows).

21. As can be seen from the varying widths of the confidence intervals in Figure 2(a)–(c), the precision of the estimated program effects varies widely across studies in our sample. The precision is essentially uncorrelated with the sample size of the evaluation (correlation $= -0.02$), suggesting that studies with larger sample sizes have more complex econometric designs that offset any potential gains in precision.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

TABLE 5. Estimated program effects on probability of employment.

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| *Effect term (omitted = short term)* | | | | | |
| Medium term | 0.035 | 0.029 | 0.045 | 0.040 | 0.032 |
| | (0.012) | (0.009) | (0.016) | (0.011) | (0.008) |
| Long term | 0.064 | 0.045 | 0.046 | 0.045 | 0.035 |
| | (0.019) | (0.015) | (0.018) | (0.017) | (0.014) |
| *Program type (omitted = training)* | | | | | |
| Job search assist. | −0.032 | −0.009 | −0.008 | 0.007 | – |
| | (0.012) | (0.020) | (0.011) | (0.021) | |
| Private subsidy | 0.042 | 0.042 | −0.009 | 0.016 | – |
| | (0.030) | (0.026) | (0.037) | (0.041) | |
| Public sector emp. | −0.065 | −0.08 | −0.056 | −0.058 | – |
| | (0.013) | (0.020) | (0.014) | (0.023) | |
| Other | 0.007 | 0.003 | 0.052 | 0.047 | – |
| | (0.027) | (0.036) | (0.026) | (0.038) | |
| *Interaction with medium term* | | | | | |
| Job search assist. | – | – | −0.037 | −0.037 | −0.028 |
| | | | (0.019) | (0.018) | (0.011) |
| Private subsidy | – | – | 0.005 | −0.012 | −0.048 |
| | | | (0.045) | (0.044) | (0.046) |
| Public sector emp. | – | – | −0.02 | −0.024 | −0.015 |
| | | | (0.027) | (0.027) | (0.015) |
| Other | – | – | −0.059 | −0.048 | −0.028 |
| | | | (0.022) | (0.021) | (0.014) |
| *Interaction with long term* | | | | | |
| Job search assist. | – | – | −0.048 | −0.03 | −0.034 |
| | | | (0.019) | (0.022) | (0.014) |
| Private subsidy | – | – | 0.153 | 0.088 | −0.061 |
| | | | (0.060) | (0.056) | (0.044) |
| Public sector emp. | – | – | −0.002 | −0.053 | −0.061 |
| | | | (0.028) | (0.039) | (0.031) |
| Other | – | – | −0.098 | −0.109 | −0.051 |
| | | | (0.030) | (0.037) | (0.015) |
| Additional controls | No | Yes | No | Yes | PPS fixed effects |

Notes: Sample size is 352 estimates. Standard errors (clustered by study) in parentheses. Models are linear regressions with the effect size as dependent variable. Coefficients of additional control variables are reported in Table 7. Model in column (5) is estimated with fixed effects controlling for 200 program participant subgroups.

We present models with and without additional controls in columns (3) and (4), and a model with dummy variables for each participant/program subgroup in column (5). In the latter specification the "main effects" for the type of program are absorbed by the PPS fixed effects, but we can still estimate the coefficients for medium and long term effects—which now measure the evolution of the program effects *for the same PPS* over different time horizons—as well as interactions of the time horizon dummies with the type of program.

Three key conclusions emerge from these more flexible specifications. First, as suggested by the patterns in Table 4, the program impacts for training programs tend

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

to rise over time, whereas the effects for job search assistance programs and other programs (which are obtained by adding the program-type/time horizon interactions to the time horizon effects in rows 1 and 2) are roughly constant.[22] Second, in the models without PPS fixed effects (columns (3) and (4)) the implied profile of impacts for private sector employment programs is relatively similar to the profile for training programs. When the PPS effects are added, however, the interactions between private sector programs and both medium term and long term horizon become relatively large and negative—similar to the interaction effects for JSA and other programs. A third conclusion is that public sector employment programs appear to be relatively ineffective at all time horizons.

We have also estimated models similar to the specifications in Table 5, but using two alternative measures of program impacts: the estimated "effect size" (the estimated effect on the employment rate of participants divided by the standard deviation of employment rates in the comparison group), and the proportional program effect (the estimated effect for participants divided by the mean employment rate of the comparison group). These specifications are reported in Online Appendix Tables A.2(a) and A.2(b), respectively, and yield very similar conclusions to the models in Table 5. Essentially, these alternative choices lead to rescaling of the coefficients of the meta-analysis models with very small changes in the relative magnitudes of different coefficients.

A limitation of the analysis in Table 5 is that estimated program effects are only available for 40% of our overall sample. To supplement these models we turn to ordered probit models for sign and significance. The first 4 columns of Table 6 present a series of OP models that are parallel to those in Table 5, but fit to our overall sample of program estimates. The specifications in columns (1) and (3) have no controls other than dummies for medium and long term horizons and the type of ALMP—in the latter case interacting the type of program with the time horizon dummies. Columns (2) and (4) report expanded specifications that add the control variables reported in Table 7. Column (5) of Table 6 repeats the specification from column (4), but fit to the subsample of 352 program estimates for which we have an estimate of the program effect on the probability of employment. The model in column (6) reproduces the specification in column (3), but adding PPS fixed effects. As in the model in the last column of Table 5, the coefficients in this specification measure the evolution of the factors determining sign and significance *within a given PPS*. Finally, column (7) presents estimates from a linear regression replacing the categorical outcome variable with values of $-1$, 0, and $+1$, also including PPS fixed effects.

22. To aid in the interpretation of the interacted estimates, Online Appendix Table A.4 presents the implied mean program effects by program type and time horizon for the models in columns (3) and (4), and the associated standard errors. Note that by construction the model in column (3) reproduces the means reported in columns (4)–(6) of Table 3(a). For the specification in column (4) of Table 5 we normalize the covariates to have mean 0 and fit the model without an intercept: thus the mean program effects are interpreted as means for a program and participant group with the mean characteristics of our sample.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

TABLE 6. Models for sign/significance of estimated program effects.

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| *Effect term (omitted = short term)* | | | | | | | |
| Medium term | 0.372 | 0.483 | 0.563 | 0.639 | 0.491 | 2.008 | 0.387 |
| | (0.088) | (0.099) | (0.130) | (0.138) | (0.145) | (0.452) | (0.158) |
| Long term | 0.597 | 0.742 | 0.901 | 1.053 | 1.03 | 2.536 | 0.457 |
| | (0.157) | (0.167) | (0.175) | (0.171) | (0.206) | (0.449) | (0.135) |
| *Program type (omitted = training)* | | | | | | | |
| Job search assist. | 0.274 | 0.286 | 0.531 | 0.532 | 0.569 | – | – |
| | (0.156) | (0.168) | (0.180) | (0.197) | (0.459) | | |
| Private subsidy | 0.139 | 0.076 | −0.04 | −0.132 | −0.166 | – | – |
| | (0.189) | (0.210) | (0.224) | (0.263) | (0.438) | | |
| Public sector emp. | −0.677 | −0.758 | −0.383 | −0.489 | −1.399 | – | – |
| | (0.219) | (0.228) | (0.276) | (0.279) | (0.496) | | |
| Other | −0.11 | −0.205 | 0.318 | 0.202 | 1.148 | – | – |
| | (0.172) | (0.184) | (0.206) | (0.236) | (0.653) | | |
| *Interaction with medium term* | | | | | | | |
| Job search assist. | – | – | −0.289 | −0.283 | −0.004 | −0.743 | −0.157 |
| | | | (0.235) | (0.249) | (0.343) | (0.618) | (0.218) |
| Private subsidy | – | – | 0.138 | 0.226 | 0.353 | −1.085 | −0.266 |
| | | | (0.289) | (0.311) | (0.486) | (1.025) | (0.289) |
| Public sector emp. | – | – | −0.645 | −0.573 | 0.051 | −1.394 | −0.229 |
| | | | (0.285) | (0.288) | (0.477) | (0.861) | (0.305) |
| Other | – | – | −0.764 | −0.705 | −0.662 | −2.04 | −0.387 |
| | | | (0.226) | (0.245) | (0.278) | (0.796) | (0.234) |
| *Interaction with long term* | | | | | | | |
| Job search assist. | – | – | −1.017 | −1.022 | −0.832 | −1.842 | −0.331 |
| | | | (0.313) | (0.294) | (0.313) | (0.775) | (0.258) |
| Private subsidy | – | – | 0.611 | 0.58 | 1.274 | −2.295 | −0.385 |
| | | | (0.375) | (0.387) | (0.798) | (1.428) | (0.350) |
| Public sector emp. | – | – | −0.643 | −0.675 | 0.131 | −2.366 | −0.450 |
| | | | (0.490) | (0.497) | (0.832) | (1.568) | (0.822) |
| Other | – | – | −0.999 | −1.021 | −1.638 | −0.735 | −0.194 |
| | | | (0.353) | (0.375) | (0.430) | (1.282) | (0.344) |
| Additional controls | No | Yes | No | Yes | Yes | PPS fixed effects | PPS fixed effects |

Notes: Sample size is 857 program estimates, except column (5), which is based on 352 estimates for which program effect on probability of employment is also available. Standard errors (clustered by study) in parentheses. Models in columns (1)–(6) are ordered probits, fit to ordinal data with value of +1 for significantly positive, 0 for insignificant, −1 for significantly negative estimate. Estimated cutpoints (2 per model) are not reported in the table. Coefficients of additional control variables are reported in Table 7. Model in column (6) is estimated with fixed effects controlling for program participant subgroups. Model in column (7) is a linear regression with fixed effects for program participant subgroups.

The OP models in Table 6 yield coefficients that are very highly correlated with the corresponding coefficients from the program effect models in Table 5, but roughly 10 times bigger in magnitude. For example, the correlation of the 14 coefficients from the specification in column (4) of Table 6 with corresponding coefficients from the

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

TABLE 7. Estimated coefficients of control variables included in models in Tables (5) and (6).

| | Program effect—OLS probit models specifications in Table 5 (columns (2), (4)) | | Sign/significance—ordered models specifications in Table 6 (columns (2), (4), (5)) | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| *Outcome of interest (omitted = probability of employment)* | | | | | |
| Earnings | – | – | −0.003 | −0.01 | – |
| | | | (0.130) | (0.132) | |
| Hazard to new job | – | – | 0.275 | 0.264 | – |
| | | | (0.211) | (0.212) | |
| Other hazard | – | – | 0.613 | 0.547 | – |
| | | | (0.275) | (0.263) | |
| Unemployment status | – | – | 0.598 | 0.591 | – |
| | | | (0.293) | (0.285) | |
| *Age of program group (omitted = mixed)* | | | | | |
| Youths (<25) | −0.031 | −0.025 | −0.368 | −0.348 | −0.518 |
| | (0.018) | (0.018) | (0.151) | (0.153) | (0.287) |
| Older (≥25) | −0.07 | −0.063 | −0.423 | −0.425 | −0.671 |
| | (0.020) | (0.020) | (0.157) | (0.160) | (0.297) |
| *Gender of program group (omitted = mixed)* | | | | | |
| Males only | 0.011 | 0.007 | −0.007 | −0.006 | −0.328 |
| | (0.022) | (0.022) | (0.149) | (0.149) | (0.266) |
| Females only | 0.047 | 0.041 | 0.064 | 0.053 | 0.000 |
| | (0.023) | (0.023) | (0.144) | (0.146) | (0.250) |
| *Country group (omitted = nordic)* | | | | | |
| Germanic | 0.056 | 0.045 | 0.250 | 0.176 | 0.910 |
| | (0.027) | (0.026) | (0.192) | (0.196) | (0.488) |
| Anglo | −0.026 | −0.028 | 0.177 | 0.14 | 1.231 |
| | (0.030) | (0.028) | (0.241) | (0.236) | (0.579) |
| East Europe | 0.022 | 0.028 | 0.131 | 0.096 | 0.618 |
| | (0.031) | (0.028) | (0.201) | (0.202) | (0.378) |
| Rest of Europe | 0.016 | 0.012 | 0.125 | 0.088 | 0.738 |
| | (0.024) | (0.023) | (0.187) | (0.189) | (0.483) |
| Latin America | 0.009 | 0.012 | 0.108 | 0.1 | 1.012 |
| | (0.053) | (0.053) | (0.338) | (0.338) | (0.826) |
| Remaining countries | 0.035 | 0.038 | −0.063 | −0.064 | 1.124 |
| | (0.035) | (0.035) | (0.281) | (0.286) | (0.529) |
| *Type of program participant (omitted = registered unemployed)* | | | | | |
| Disadvantaged | 0.018 | 0.013 | 0.542 | 0.527 | 0.356 |
| | (0.036) | (0.036) | (0.228) | (0.228) | (0.623) |
| Long-term unemployed | 0.083 | 0.08 | 0.388 | 0.404 | 0.392 |
| | (0.032) | (0.031) | (0.181) | (0.179) | (0.332) |
| Program duration— | −0.029 | −0.023 | −0.135 | −0.122 | −0.55 |
| Dummy if >9 months | (0.016) | (0.016) | (0.179) | (0.177) | (0.232) |
| Randomized experimental | −0.009 | −0.008 | −0.065 | −0.095 | −0.314 |
| Design | (0.020) | (0.019) | (0.170) | (0.170) | (0.332) |
| Square root of sample size | −0.003 | 0.001 | 0.159 | 0.098 | 0.484 |
| | (0.042) | (0.037) | (0.184) | (0.191) | (0.706) |
| Published article | −0.024 | −0.026 | −0.203 | −0.213 | −0.41 |
| | (0.017) | (0.017) | (0.133) | (0.132) | (0.254) |
| Citations rank index | −0.001 | −0.001 | 0.007 | 0.005 | −0.005 |
| | (0.002) | (0.001) | (0.012) | (0.012) | (0.024) |
| *R*-squared/log likelihood | 0.36 | 0.40 | −765 | −752 | −283 |

Notes: Standard errors (clustered by study) in parentheses. Coefficients in columns (1) and (2) are from models reported in columns (2) and (4) of Table 5. Coefficients in columns (3)–(5) are from models reported in columns (2), (4), and (5) of Table 6. See notes to Tables 5 and 6 for more information.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

specification in column (4) of Table 5 is 0.84.[23] In particular the OP models confirm that the impacts of job search assistance and other programs tend to fade relative to the impacts of training programs, and that public sector employment programs are relatively ineffective at all time horizons, regardless of how the outcomes are measured in the evaluation.[24]

The models including PPS fixed effects in columns (6) (ordered probit) and (7) (linear regression) of Table 6 also imply the same qualitative findings, indicating that within a given PPS the estimated program effect becomes more positive in the longer run. The coefficients of the linear regression model are scaled by a factor of approximately five relative to the ordered probit specification.

### *4.3. Participant and Study Characteristics in the Basic Models*

The estimated coefficients for the extra control variables included in the models in columns (2), (4) of Table 5 and columns (2), (4), and (5) of Table 6 are reported in Table 7. The coefficient estimates from the two models for the effects on the probability of employment (columns (1), (2)) are quite similar and suggest that the impact of ALMPs varies systematically with the type of participant (with larger effects for the long term unemployed), their age group (more negative impacts for older and younger participants), and their gender (larger effects for females). The estimated program effects are also somewhat larger for studies estimated on German, Austrian, or Swiss data, but there are no large or significant differences across the other country groups.

The coefficients from the OP models (columns (3)–(4)) confirm most of these conclusions about the differential impacts of ALMPs across different participant groups and different countries.[25] In particular, the OLS models for the program effects on the probability of employment and the OP models for sign and significance show smaller impacts for young participants and older participants, relative to the impacts on mixed age groups, and larger impacts for long-term unemployed participants. The OP models fit to the overall sample (columns (3), (4)) also point to a larger positive impact for disadvantaged participants relative to UI recipients, whereas the program effect models and the OP models fit to the program effect subsample (column (5)) yield an insignificant coefficient, arguably due to the small number of studies that focus on this group.[26]

23. The regression model is: OP-coefficient $= -0.02 + 10.57 \times$ effect-coefficient, $R$-squared $= 0.70$.

24. We also fit two simpler probit models for the events of reporting a positive and significant or negative and significant estimate, reported in Online Appendix Table A.3. As would be expected if the ordered probit specification is correct, the coefficients from the model for a significantly positive effect are quite close to the OP coefficients, while the coefficients from the model for a significantly negative effect are close in magnitude but opposite in sign.

25. The correlation between the coefficients in columns (2) and (4) of Table 7 is 0.69.

26. A potentially relevant dimension of program effectiveness concerns the time ALMP participants have spent in unemployment before entering the program. Biewen et al. (2014) investigate this issue using

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

One notable difference between the program effect models and the OP models concerns the relative impact of ALMPs on female participants. In the OLS program effect models the estimated coefficients for female participants are around 0.04–0.05 in magnitude, and statistically significant at conventional levels (with $t$ statistics around 2). In the OP models, by comparison, the corresponding coefficients are relatively small in magnitude and far from significant. Further investigation reveals that this divergence is driven by the upper tail of program effect estimates for female participants (see Online Appendix Figure A.4), and in particular by the relatively large estimated effects for female PPSs that show a significant positive effect.[27] This upper tail does not appear to be driven by a few outliers, but instead reflects a systematically higher probability of estimating a large positive effect when the participant group is limited to females.[28]

An interesting aspect of the OP models is the pattern of coefficients associated with the choice of dependent variable, reported in the top rows of Table 7. These coefficients show that studies modeling the hazard rate of exiting the benefit system or the probability of unemployment are significantly more likely to report positive findings than studies modeling employment (the omitted category) or earnings.[29]

The models summarized in Table 7 also control for the duration of the program, using a simple dummy variable for whether the program lasted longer than 9 months. Program duration can be seen as a rough proxy for the cost of the program, as information on cost effectiveness is very sparse in most of the surveyed studies. Our estimates do not indicate any systematic advantage for longer programs—indeed across all the specifications the coefficient is negative, though statistically insignificant.

three strata to estimate treatment effects: 1–3 months, 4–6 months, and 7–12 months of unemployment, respectively. For longer training programs (mean duration of 226 days) the estimated short-term impacts and standard errors for the three strata are 0.00 (0.04), 0.00 (0.04), 0.08 (0.40) for males, and 0.06 (0.03), −0.01 (0.045), −0.04 (0.02) for females; medium-term impacts are 0.05 (0.03), 0.07 (0.04), 0.09 (0.045) for males, and 0.06 (0.03), 0.11 (0.05), 0.09 (045) for females. These results show some indication that program participants from strata with longer elapsed unemployment durations benefit more than other groups. This result is in line with the findings from our stratification of program intake group into short-term unemployed ("UI recipients"), long-term unemployed, and participants without benefit entitlement ("disadvantaged"). A more detailed investigation of this issue within the meta-analysis framework is not possible, since the primary studies rarely report information on elapsed time in unemployment before program start.

27. The median and 75th percentiles of the effect size distribution for female participant groups, conditional on a positive impact, are 0.25 and 0.46, respectively. By comparison, the corresponding statistics for the combined male and mixed gender participant groups are 0.15 and 0.27.

28. We also estimated separate program effect models for different types of participants—those from the regular UI system versus long term unemployed or disadvantaged groups. We found a significant positive coefficient for female participants in the models for both UI recipients and the long term unemployed.

29. Estimates from interacted models that allow different effects of the dependent variable at different time horizons (not reported in the table) show that the positive coefficient associated with the use of exit hazards is largely confined to short term impacts.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

### *4.4. Randomized versus Nonexperimental Designs*

A longstanding concern in the ALMP literature (e.g., Ashenfelter 1987) is the unreliability of nonexperimental estimators. This concern led to a series of large scale experiments in the United States designed to test job training programs (Bloom et al. 1997; Schochet, Burghardt, and Glazerman 2001) and a large literature comparing experimental and nonexperimental estimates for the same program (e.g., Lalonde 1986; Smith and Todd 2005).

One simple way to assess these concerns is to compare the magnitudes of the estimated effects from papers based on experimental designs to those from nonexperimental designs. To do this, we include a simple indicator for an experimental design in the models in Table 7. In the program effect models (columns (1) and (2)) the coefficient of the dummy is very small in magnitude (less than 1.0 ppt.) and insignificantly different from zero. Likewise, in the main OP models (columns (3)–(4)) the coefficient is small in magnitude. It is a little larger in magnitude when the OP models are restricted to the set of studies with estimated program effects (column (5)) but relatively imprecise, perhaps reflecting the relatively large number of parameters included in the model relative to the sample size or the uneven distribution of experimental evaluations across program types. We conclude from this analysis that there is little or no evidence that results from experimentally-based designs in the recent ALMP literature are "less positive" or "less significant" than results from nonexperimental designs.

### *4.5. Publication Bias and p-Hacking*

A related concern, widely discussed in the meta-analysis literature (e.g., Rothstein, Sutton, and Borenstein 2005) is that the set of estimated program impacts in the available literature contain a systematic positive bias, either because analysts only write up and circulate studies that show a positive effect (so-called *file drawer bias*) or because they choose specifications that tend to yield positive and significant effects (so-called *p-hacking*).

A standard way to look for evidence of publication bias is to examine funnel plots of the relationship between the estimated program effects and their precision (Sutton et al. 2000). Figure 3(a)–(c) present these plots for the program effects for employment in our sample, restricting attention (as in Figure 2(a)–(c)) to estimates that have a corresponding sampling error available. For reference, we also show the boundaries of the "$t = 2$" relationship in each graph.[30] Contrary to the inverted funnel pattern typically uncovered in the meta-analysis literature (e.g., Doucouliagos and Stanley 2009; Havránek 2015; Wolfson and Belman 2015), there is a lot of dispersion in the estimated program effects at all three time horizons, even among studies with

30. Since $t = P^{1/2}\, b$, the "$t = 2$" relationship is $P = 4/b^2$ which is a pair of hyperbolas centered around the y axis in a funnel plot.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

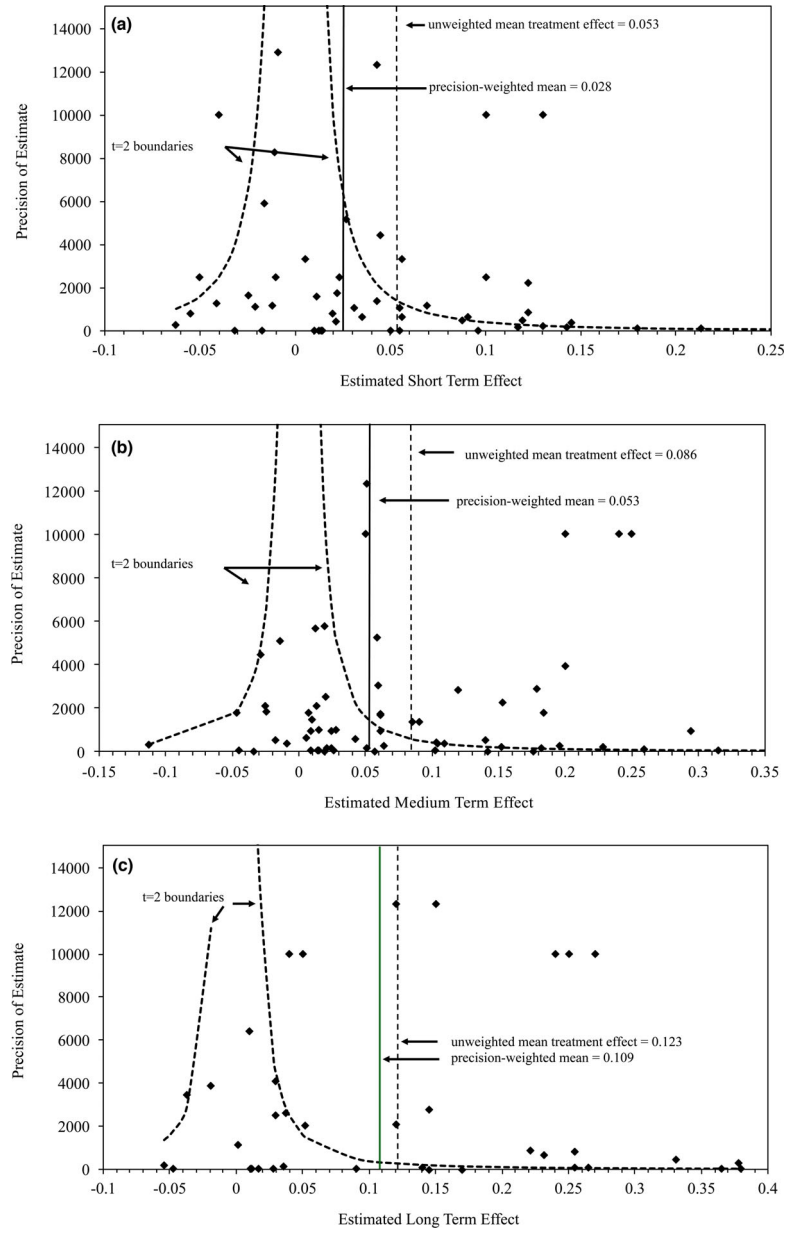


FIGURE 3. (a) Funnel plot of short term estimates. Diamonds represent precision of estimated short term treatment effect for a program/participant (PPS) subgroup, graphed against the estimated treatment effect. Graph shows 56 estimates—2 large positive estimates are not shown for clarity. (b) Funnel plot of medium term estimates. Diamonds represent precision of estimated medium term treatment effect for a program/participant subgroup (PPS), graphed against the estimated treatment effect. Graph shows 69 estimates—3 large positive estimates are not shown for clarity. (c) Funnel plot of long term estimates. Diamonds represent precision of estimated long term treatment effect for a program/participant (PPS) subgroup, graphed against the estimated treatment effect. Graph shows 39 estimates.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

TABLE 8. Tests for publication bias (funnel asymmetry tests).

| | Control for time horizon only | Basic controls (column (2) of Table 5) | Interacted controls (column (4) of Table 5) | PPS fixed effects (column (5) of Table 5) |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| *Estimated coefficient of sampling error of estimated program effect* | | | | |
| Unweighted OLS | 0.007 | 0.009 | 0.009 | 0.001 |
| | (0.005) | (0.008) | (0.008) | (0.014) |
| Precision-weighted | 0.022 | 0.035 | 0.032 | −0.005 |
| | (0.031) | (0.024) | (0.020) | (0.019) |

Notes: Entry in each row and column corresponds to estimated coefficient of sampling error of estimated program effect from a different specification. Models in column (1) include only a constant and dummies for medium term and long term time horizon as additional controls; models in columns (2), (3), and (4) include same controls included in specifications in columns (2), (4), and (5) of Table 5, respectively. Unweighted models are fit by OLS. Precision-weighted models are fit by weighted least squares using Winsorized inverse sampling variance of estimated program effect as weight. Weight is Winsorized at 10th and 90th percentiles, respectively, corresponding to values of 19.2 and 10,000, respectively. Standard errors, clustered by study, in parentheses.

high levels of precision, suggesting that the variation in the estimates is not just a result of sampling error.

A more formal test for publication bias is to regress the estimated program effect from a given study and specification on the associated sampling error of the estimate and other potential control variables. Using the notation of Section 4.1, the regression model is

$$b = X\alpha + \theta P^{-1/2} + \nu \tag{4}$$

where $\nu$ represents a residual. The estimate of $\theta$ is interpreted as a test for asymmetry in the funnel plot relationship between the estimated program effects and their precision: if the sample contains more imprecisely estimated large positive effects than large negative effects, $\theta$ will be positive. Stanley (2008) suggests that the model be estimated by weighted least squares, using the precision of each estimate (i.e., its inverse sampling variance) as a weight. If sampling error is the only source of residual variation in (4) this will lead to efficient estimates.

Estimation results for this model are presented in Table 8. We present estimates from four specifications estimated by unweighted OLS and precision-weighted least squares.[31] The results give no indication of publication bias. The unweighted estimate in column (3), for example, which is based on a model that includes our richest set of controls, implies that a study with a 1 percentage point larger standard error for the estimated program effect on the probability of employment will have about a

31. The estimated precision of the estimates in our sample ranges from 0.026 to over 50,000. To stabilize the estimates we Winsorize the precision weights at the 10th and 90th percentiles.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

0.01 ppt. larger estimated program effect. The corresponding weighted estimate suggests a slightly larger 0.03 ppt. larger estimate, but is less precise.

We suspect that there are at least two explanations for the apparent lack of publication bias in the ALMP literature. The first is that ALMP evaluations are often conducted in close cooperation with the agencies that operate the programs. In such settings researchers cannot simply shelve papers that show small or "wrong signed" impacts. They may also find it hard to choose among specifications to obtain a more positive program effect. A second factor is that referees and other researchers have no strong presumption that ALMP's necessarily "work", or that a finding of a negative or insignificant effect is uninteresting, since many important papers in the field (e.g., Lalonde 1986; Heckman and Hotz 1989) report small or negative impacts of ALMPs.

Our models also control for two other features of a study that may be informative about the presence of publication bias: whether it was published, and the number of citations it received (measured from a Google Scholar search in Spring 2015).[32] The coefficients associated with both variables (shown in Table 7) are small and insignificant across all specifications, confirming that there is no tendency for more positive studies to be published or to be more highly cited.

### *4.6. Are Some Programs Better (or Worse) for Different Participant Groups?*

A longstanding question in the ALMP literature is whether certain participant groups are "better matched" to specific types of programs (for an analysis in the German context see Biewen et al. 2007). We address this in Table 9, which presents separate models for the program effects from different types of ALMPs.

As a benchmark column (1) presents a baseline specification fit to all 5 program types, with dummies for the program types (not reported) and controls for the intake group, the gender group, and the age group.[33] The (omitted) base group is comprised of mixed gender and age groups from the regular UI rolls. In this pooled specification the estimated effects for females and long term unemployed participants are significantly positive, whereas the coefficient for older participants is significantly negative, and the coefficient for young participants is negative and marginally significant.

Columns (2)–(6) report estimates for the same specification (minus the controls for the type of program) fit separately to the estimated effects for each of the 5 program types. Comparisons across these models suggest that long-term unemployed participants benefit relatively more from "human capital" programs (i.e., training

32. To account for lags in the citation process we model citations as the rank within the distribution of citations for papers written in the same year.

33. This is a simplified version of the specification reported in column (2) of Table 5 and column (1) of Table 7.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

TABLE 9. Comparison of estimated program effects from different program types on different participant groups.

| | All program types (1) | Training (2) | Job search assistance (3) | Private sector job/ subsidy (4) | Public sector employ-ment (5) | Other (6) |
|---|---|---|---|---|---|---|
| Number estimates | 352 | 217 | 36 | 46 | 32 | 21 |
| Number of studies | 83 | 51 | 15 | 19 | 14 | 8 |
| Mean program effect (×100) | 4.52 | 4.71 | 1.48 | 9.91 | −1.83 | 5.65 |
| Constant | 0.024 | 0.020 | −0.003 | −0.053 | −0.004 | 0.106 |
| | (0.015) | (0.014) | (0.017) | (0.043) | (0.027) | (0.022) |
| Medium term | 0.032 | 0.041 | 0.017 | 0.028 | 0.020 | 0.004 |
| | (0.009) | (0.010) | (0.014) | (0.045) | (0.018) | (0.019) |
| Long term | 0.054 | 0.055 | 0.005 | 0.133 | 0.009 | −0.011 |
| | (0.016) | (0.016) | (0.012) | (0.068) | (0.023) | (0.014) |
| *Intake group (base=regular UI recipients)* | | | | | | |
| Disadvantaged | −0.002 | −0.020 | 0.037 | 0.053 | (omitted) | 0.030 |
| | (0.019) | (0.020) | (0.017) | (0.026) | | (0.030) |
| Long term unemployment | 0.072 | 0.122 | 0.024 | 0.088 | 0.029 | −0.109 |
| | (0.032) | (0.059) | (0.018) | (0.038) | (0.030) | (0.019) |
| *Gender group (base=mixed)* | | | | | | |
| Male | 0.019 | 0.028 | (omitted) | 0.110 | −0.048 | −0.022 |
| | (0.020) | (0.025) | | (0.069) | (0.031) | (0.023) |
| Female | 0.054 | 0.058 | (omitted) | 0.163 | −0.024 | −0.111 |
| | (0.022) | (0.028) | | (0.050) | (0.030) | (0.025) |
| *Age group (base=mixed)* | | | | | | |
| Youth | −0.037 | −0.030 | 0.009 | 0.034 | −0.059 | (omitted) |
| | (0.018) | (0.021) | (0.011) | (0.037) | (0.030) | |
| Older participants | −0.048 | −0.059 | 0.016 | −0.094 | −0.047 | 0.046 |
| | (0.019) | (0.024) | (0.021) | (0.053) | (0.038) | (0.002) |
| Controls for program type[a] | Yes | No | No | No | No | No |

Notes: Standard errors, clustered by study, in parenthesis. See note to Table 5.

a. Four dummies for different types of programs included in column (1) only.

and private sector employment), and relatively less from "work first" programs (i.e., job search and other programs). In contrast, disadvantaged participants appear to benefit more from work first programs and less from human capital programs. Female participants also appear to benefit relatively more from training and private sector subsidy programs, whereas the relative effects for youths and older participants are not much different across the program types.

Overall these results suggest that there may be potential gains to matching specific participant groups to specific types of programs, though the small sample sizes for most of the program types must be noted. Attempts to expand the power of the analysis by using OP models for the sign and significance of the program estimates lead to generally similar conclusions as the program effect models reported in Table 9 with only modest gains in precision.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

### *4.7. Effects of Cyclical Conditions*

Another longstanding question in the ALMP literature is whether programs are more (or less) effective in different cyclical environments.[34] One view is that active programs are *less* effective in a depressed labor market because participants have to compete with other, more advantaged workers for a limited set of jobs. An alternative view is that ALMPs are *more* effective in weak labor markets because employers become more selective in a slack market, increasing the value of an intervention that makes workers more job-ready.

Three previous studies have investigated ALMP effectiveness over the business cycle. Kluve (2010) uses between-country variation in a small European meta data set, whereas Lechner and Wunsch (2009) and Forslund, Fredriksson, and Vikström (2011) analyze programs in Germany and Sweden, respectively. All three studies suggest a positive correlation between ALMP effectiveness and the unemployment rate.

To provide some new evidence we added two alternative contextual variables to our analysis, representing the average growth rate of GDP and the average unemployment rate during the years the treatment group participated in the program. Since growth rates and unemployment rates vary widely across countries, we also introduced a set of country dummies that absorb any permanent differences in labor market conditions across countries. The effect of these dummies is interesting in its own right because the shares of different program types and participant groups also vary widely across countries, leading to the possibility of bias in the measured effects of program types and participant groups if there are unobserved country specific factors that affect the average success of ALMPs in different countries.

The results of our analysis are summarized in Table 10. For reference column (1) presents a benchmark specification identical to the simplified model in column (1) of Table 9. Column (2) presents the same specification with the addition of 37 country dummies. The addition of these dummies leads to some modest but interesting changes in the estimated coefficients in the meta-analysis model. Most notably, the coefficients associated with job search assistance (JSA) and "other" programs both become more negative, indicating that these programs tend to be more widely used in countries where all forms of ALMPs are relatively successful.

Column (3) presents a model that includes the control for average GPD growth rate during the program period. The coefficient is negative and marginally significant ($t = 1.7$) providing suggestive evidence that ALMPs work better in recessionary markets. A model that controls for the average unemployment rate shows the same tendency (coefficient = 0.006, standard error = 0.007) though the effect is less precise.

A concern with the specification in column (3) is that the average number of program estimates per country is small (many countries have only 2 or 3 estimates) leading to potential overfitting. To address this, we estimated the models in columns

34. A related question is whether program externalities are bigger or smaller in weak or strong labor markets. This is addressed in the experiment conducted by Crépon et al. (2013).

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

TABLE 10. Impacts of macro conditions on the effectiveness of ALMP's.

| | All available program effect estimates | | | Denmark, France, Germany, and US only | | |
|---|---|---|---|---|---|---|
| | Baseline (1) | +Country effects (2) | +GDP growth (3) | Baseline (4) | +GDP growth (5) | +Unemp. rate (6) |
| Medium term | 0.032 | 0.030 | 0.028 | 0.043 | 0.034 | 0.040 |
| | (0.009) | (0.009) | (0.009) | (0.009) | (0.008) | (0.009) |
| Long term | 0.054 | 0.046 | 0.040 | 0.056 | 0.031 | 0.048 |
| | (0.016) | (0.017) | (0.015) | (0.020) | (0.014) | (0.020) |
| GDP growth rate (%) | – | – | −0.010 | – | −0.032 | 0.034 |
| (unemp. rate in column (6)) | | | (0.006) | | (0.008) | (0.011) |
| *Program type (base=training)* | | | | | | |
| Job search assistance | −0.033 | −0.053 | −0.056 | −0.076 | −0.103 | 0.010 |
| | (0.017) | (0.028) | (0.027) | (0.034) | (0.023) | (0.069) |
| Private sector job/subsidy | 0.031 | 0.027 | 0.019 | 0.020 | −0.002 | 0.012 |
| | (0.026) | (0.028) | (0.028) | (0.029) | (0.024) | (0.031) |
| Public sector employment | −0.079 | −0.074 | −0.069 | −0.096 | −0.073 | −0.102 |
| | (0.019) | (0.027) | (0.025) | (0.029) | (0.024) | (0.025) |
| Other programs | −0.015 | −0.045 | −0.066 | −0.096 | −0.188 | −0.104 |
| | (0.033) | (0.027) | (0.033) | (0.034) | (0.046) | (0.050) |
| *Intake group (base=regular UI recipients)* | | | | | | |
| Disadvantaged | −0.002 | 0.000 | 0.010 | 0.046 | 0.106 | 0.050 |
| | (0.019) | (0.034) | (0.033) | (0.028) | (0.035) | (0.027) |
| Long term unemployed | 0.072 | 0.098 | 0.095 | 0.112 | 0.109 | 0.092 |
| | (0.032) | (0.033) | (0.031) | (0.034) | (0.027) | (0.033) |
| *Gender group (base=mixed)* | | | | | | |
| Female | 0.019 | 0.048 | 0.053 | 0.066 | 0.081 | 0.052 |
| | (0.020) | (0.025) | (0.024) | (0.026) | (0.022) | (0.022) |
| Male | 0.054 | 0.086 | 0.092 | 0.094 | 0.111 | 0.081 |
| | (0.022) | (0.030) | (0.030) | (0.033) | (0.030) | (0.029) |
| *Age group (base=mixed)* | | | | | | |
| Youth | −0.037 | −0.026 | −0.026 | −0.024 | −0.057 | −0.040 |
| | (0.018) | (0.018) | (0.020) | (0.023) | (0.020) | (0.051) |
| Older participants | −0.048 | −0.055 | −0.062 | −0.073 | −0.097 | −0.066 |
| | (0.019) | (0.024) | (0.025) | (0.028) | (0.023) | (0.023) |
| Country dummies | No | Yes | Yes | Yes | Yes | Yes |

Notes: Standard errors, clustered by study, in parenthesis. Models in columns (1)–(3) are fit to 352 program estimates from 83 studies, with mean of dependent variable = 0.0452. Models in columns (4) and (5) are fit to 200 program estimates from Denmark, France, Germany, and the United States from 38 studies, with mean of dependent variable = 0.0423. Model in columns 6 is fit to 181 program estimates from the same four countries from 34 studies, with mean of dependent variable = 0.0441.

(4)–(6), using only data from the four countries that account for the largest numbers of program estimates—Denmark (17 estimates), France (20 estimates), Germany (147 estimates) and the United States (16 estimates). As shown in column (4), our baseline specification yields coefficient estimates that are quite similar to the estimates from

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

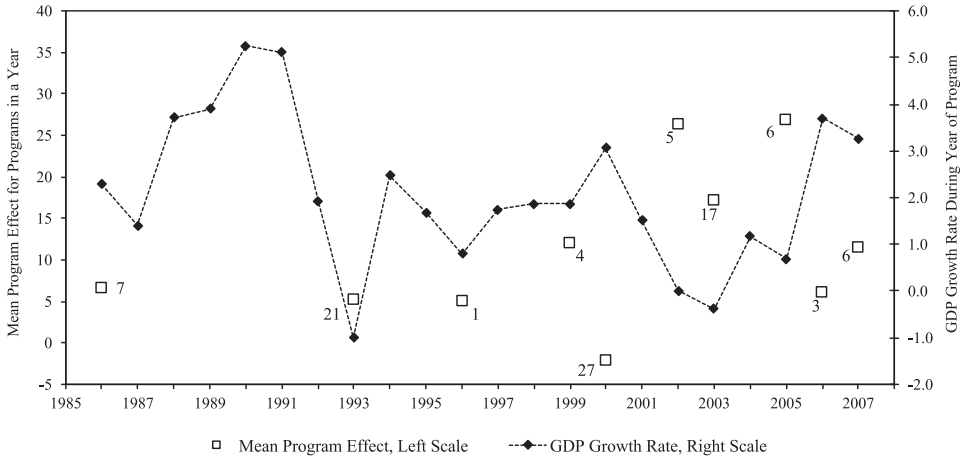


FIGURE 4. Mean program effects and GDP growth rate—German ALMP evaluations. Mean program effect on the employment rate of ALMP participants (in percentage points) is plotted for programs offered in different years, with number of program estimates in that year. GDP growth rate is for the first year the program was operated.

the entire sample, though the relative impacts of JSA and other programs are more negative in these 4 countries.

Columns (5) and (6) present models that add the average GDP growth rate and the average unemployment rate, respectively, to this baseline model. These specifications suggest relatively important cyclical effects on ALMP effectiveness. For example, comparing two similar programs operating in labor markets with a 3 percentage point gap in growth rates, the program in the slower growth environment would be expected to have a 0.1 larger program effect.

To illustrate the variation driving the results in columns (5) and (6), Figure 4 plots the annual GDP growth rate in Germany along with the mean program effects in different years, grouping program estimates by the year that the program was first operated. We also show the number of program estimates for each year, which varies substantially over time. The figure shows the counter-cyclical pattern of program effects is mainly driven by large positive effects of programs implemented in the early 2000s during a period of slow growth.

Although policy makers have to decide whether to increase spending for active programs and enroll more participants knowing only *current* business cycle conditions, it is also of interest how labor market conditions at the time of program completion are related to the program effects. In Online Appendix Table A.5 we estimate alternative specifications that control for the GDP growth rate and unemployment rate at the beginning and after the end of the program period. These estimates suggest that ALMP programs tend to be particularly successful if participants are enrolled in a program during a downturn and exit the program during a period of favorable economic conditions.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

Although the evidence in Table 10 suggests a countercyclical pattern of program effectiveness, it is worth emphasizing that the explanation for this pattern is less clear. It is possible that the value of a given program is higher in a recessionary environment. It is also possible, however, that the characteristics of ALMP participants, or of the programs themselves, change in a way that contributes to a more positive impact in a slow-growth/high-unemployment environment.

## 5. Summary and Conclusions

We have assembled and analyzed a new sample of impact estimates from 207 studies of active labor market policies. Building on our earlier study (CKW), we argue that it is important to distinguish between impacts at various time horizons since completion of the program, and to consider how the time profile of impacts varies by the type of ALMP. We also study the importance of participant heterogeneity, and look for evidence that specific subgroups may benefit more or less from particular types of programs. Finally, we study how the state of the labor market affects the measured effectiveness of ALMPs.

With regard to the impacts of different types of ALMPs, we find that the time profiles of "work first" style programs that offer job search assistance or incentives to enter work quickly differ from the profiles of "human capital" style training programs and public sector employment programs. Human capital programs have small (or in some cases even negative) short term impacts, coupled with larger impacts in the medium or longer run (2–3 years after completion of the program), whereas the impacts from work first programs are more stable. We also confirm that public sector employment programs have negligible, or even negative program impacts at all time horizons.

With regard to different participant groups, we find that female participants and those drawn from the pool of long term unemployed tend to have larger program effects than other groups. In contrast, the program estimates for youths and older workers are typically less positive than for other groups. We also find indications of potential gains to matching different participant groups to specific programs, with evidence that work first programs are relatively more successful for disadvantaged participants, whereas human capital programs are more successful for the long term unemployed.

With regard to the state of the labor market, we find that ALMPs tend to have larger impacts in periods of slow growth and higher unemployment. In particular, we find a relatively large cyclical component in the program estimates from four countries that account for one-half of our sample. We also find suggestive evidence that human capital programs are more cyclically sensitive than work first programs.

Our findings on the relative efficacy of human capital programs for long term unemployed, and on the larger impacts of these programs in recessionary environments, point to a potentially important policy lesson. As noted by Krueger, Judd, and Cho (2014) and Kroft et al. (2016), the number of long term unemployed rises rapidly as a recession persists. This group has a high probability of leaving the labor force,

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

risking permanent losses in the productive capacity of the economy. One policy response is countercyclical job training programs and private employment subsidies, which are particularly effective for the longer-term unemployed in a recessionary climate.

Methodologically, we find a number of interesting patterns in the recent ALMP literature. We find that the estimated impacts derived from randomized controlled trials, which account for one-fifth of our sample, are not much different on average from the nonexperimental estimates. We also find no evidence of "publication bias" in the relationship between the magnitude of the point estimates from different studies and their corresponding precision. We do find that the choice of outcome variable used in the evaluation matters, with a tendency toward more positive short term impact estimates from studies that model the time to first job than from studies that model the probability of employment or the level of earnings.

Finally, we conclude that meta-analytic models based on the sign and significance of the program impacts lead to very similar conclusions as models based on program effects. We argue that this arises because much of the variation in the sign and significance of estimated impacts across studies in the ALMP literature is driven by variation in estimated program effects, rather than by variation in the corresponding sampling errors of the estimates.

## References

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

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

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

Bergemann, Annette and Gerard J. van den Berg (2008). "Active Labor Market Policy Effects for Women in Europe—A Survey." *Annales d'Economie et de Statistique*, 91/92, 385–408.

Biewen, Martin, Bernd Fitzenberger, Aderonke Osikominu, and Marie Waller (2014). "The Effectiveness of Public Sponsored Training Revisited: The Importance of Data and Methodological Choices." *Journal of Labor Economics*, 32, 837–897.

Biewen, Martin, Bernd Fitzenberger, Aderonke Osikominu, and Marie Waller (2007). "Which Program for Whom? Evidence on the Comparative Effectiveness of Public Sponsored Training Programs in Germany." IZA Discussion Paper No. 2885.

Blank, Rebecca, David Card, and Philip K. Robins (2000). "Financial Incentives for Increasing Work and Income Among Low-Income Families." In *Finding Jobs: Work and Welfare Reform*, edited by Rebecca M. Blank and David Card. Russell Sage Foundation, New York.

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos (1997). "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study." *Journal of Human Resources*, 32, 549–576.

Card, David and Dean R. Hyslop (2005). "Estimating the Effect of a Time-Limited Earnings Subsidy for Welfare Recipients." *Econometrica*, 73, 1723–1770.

Card, David, Jochen Kluve, and Andrea Weber (2010). "Active Labour Market Policy Evaluations: A Meta-analysis." *Economic Journal*, 120, F452–F477.

Cho, Yoonyong and Maddalena Honorati (2014). "Entrepreneurship Programs in Developing Countries: A Meta Regression Analysis." *Labour Economics*, 28, 110–130.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora (2013). "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment." *Quarterly Journal of Economics*, 128, 531–580.

Doucoulias, Hristos and Tom D. Stanley (2009). "Publication Selection Bias in Minimum Wage Research? A Meta-Regression Analysis." *British Journal of Industrial Relations*, 47, 406–428.

Filges, Trine, Geir Smedslund, Anne-Sofie Due Knudsen, and Anne-Marie Klint Jørgensen (2015). "Active Labour Market Programme Participation for Unemployment Insurance Recipients: A Systematic Review." *Campbell Systematic Reviews*, 2015:2, doi: 10.4073/csr.2015.2.

Forslund, Anders, Peter Fredriksson, and Johan Vikström (2011). "What Active Labor Market Policy Works in a Recession?" *Nordic Economic Policy Review,* 1, 171–207.

Friedlander, Daniel, David H. Greenberg, and Philip K. Robins (1997). "Evaluating Government Training Programs for the Economically Disadvantaged." *Journal of Economic Literature*, 25, 1809–1855

Gladden, Tricia and Christopher Taber (2000). "Wage Progression Among Less Skilled Workers." In *Finding Work: Jobs and Welfare Reform*, edited by Rebecca M. Blank and David Card. Russell Sage Foundation, New York.

Greenberg, David H., Charles Michalopoulos, and Philip K. Robins (2003). "A Meta-Analysis of Government-Sponsored Training Programs." *Industrial and Labor Relations Review*, 57, 31–53.

Grimm, Michael and Anna Luisa Paffhausen (2015). "Do Interventions Targeted at Micro-Entrepreneurs and Small and Medium-Sized Firms Create Jobs? A Systematic Review of the Evidence for Low and Middle Income Countries." *Labour Economics*, 32, 67–85.

Ham, John C. and Robert J. Lalonde (1996). "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training." *Econometrica*, 64, 175–205.

Havránek, Tomáš (2015). "Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting." *Journal of the European Economic Association*, 13, 1180–1204.

Heckman, James J. and V. Joseph Hotz (1989). "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American statistical Association*, 84, 862–874.

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

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith (1999). "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics*, Vol. 3a, Orley Ashenfelter and David Card. Elsevier, Amsterdam.

Hedges, Larry V. and Ingram Olkin (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando.

Ibarrarán, P. and D. Rosas (2009). "Evaluating the Impact of Job Training Programs in Latin America: Evidence from IDB funded operations." *Journal of Development Effectiveness*, 1, 195–216.

Kluve, Jochen (2010). "The Effectiveness of European Active Labor Market Programs." *Labour Economics*, 17, 904–918.

Kroft, Kory, Fabian Lange, Matthew J. Notowidigdo, and Lawrence F. Katz (2016). "Long-term Unemployment and the Great Recession: The Role of Composition, Duration Dependence, and Non-Participation." *Journal of Labor Economics*, 34, S7–S54.

Krueger, Alan B., Cramer Judd, and David Cho (2014). "Are the Long-Term Unemployed on the Margins of the Labor Market?" *Brookings Papers on Economics Activity*, 2, 229–299.

Lalonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, 74(4), 604–620.

Lalonde, Robert J. (2003). "Employment and Training Programs." In *Means Tested Transfer Programs in the United States*, Robert A. Moffit. University of Chicago Press, Chicago.

Lechner, Michael and Conny Wunsch (2009). "Are Training Programs More Effective when Unemployment is High?" *Journal of Labor Economics*, 27, 653–692.

Martin, John P. (2014). "Activation and Active Labour Market Policies in OECD Countries: Stylized Facts and Evidence on Their Effectiveness." IZA Policy Paper No. 84, Bonn, Germany.

Meyer, Bruce D. (1995). "Lessons from the U.S. Unemployment Insurance Experiments." *Journal of Economic Literature*, 33, 91–131.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019

Mincer, Jacob (1974). *Schooling, Experience and Earnings*. National Bureau of Economic Research, New York.

Roberts, Colin D. and Tom D. Stanley, editors (2005). *Meta-Regression Analysis: Issues of Publication Bias in Economics*. Wiley, New York.

Rothstein, Hannah R., Alexander J. Sutton, and Michael Borenstein (2005). *Publication Bias in Meta Analysis—Prevention, Assessment and Adjustments*. Wiley, New York.

Schochet, Peter Z., John Burghardt, and Steven Glazerman (2001). *National Job Corps Study: The Impacts of Job Corps on Participants' Employment and Related Outcomes*. Mathematica Policy Research, Princeton, NJ.

Sianesi, Barbara (2004). "An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s." *Review of Economics and Statistics*, 86, 133–155.

Smith, Jeffrey A. and Petra Todd (2005). "Does Matching Overcome LaLonde's Critique of Nonexperimental Methods?" *Journal of Econometrics*, 125, 305–353.

Stanley, Tom D. (2008). "Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection." *Oxford Bulletin of Economics and Statistics*, 70, 102–127.

Stanley, Tom D. and Hristos Doucouliagos (2012). *Meta-Regression Analysis in Economics and Business*. Routledge, Milton Park, UK.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge (2015). "What Are We Weighting For?" *Journal of Human Resources*, 50, 301–316.

Sutton, A. J., S. J. Duval, R. L. Tweedie, K. R. Abrams, and D. R. Jones (2000). "Empirical Assessment of Effect of Publication Bias on Meta-analyses." *British Medical Journal*, 320, 1574–1577.

United States Department of Labor (2016). "Labor Force Statistics from the Current Population Survey 2015. http://www.bls.gov/cps Accessed on 24 August, 2017.

Wolfson, Paul J. and Dale Belman (2015). "15 Years of Research on U.S. Employment and the Minimum Wage." Tuck School of Business Working Paper No. 2705499. Available at SSRN: https://ssrn.com/abstract=2705499 or http://dx.doi.org/10.2139/ssrn.2705499

## Supplementary Data

Supplementary data are available at *JEEA* online.

Downloaded from https://academic.oup.com/jeea/article-abstract/16/3/894/4430618 by guest on 10 September 2019