BIS Evaluation Summary and Peer Review

The BIS Expert Peer Review Group for Evaluation reviews all BIS impact evaluation publications, and provides an independent assessment of the methodological quality of the evaluation.

In addition to their assessment, the reviewers also provide helpful comments and suggestions for improving the clarity and reporting of the analysis. Many of the reviewers' suggestions are implemented by the authors for the final version of the publication.

Title: Impact Of Skills And Training Interventions On The Unemployed: Phase II Report	
Programme evaluated: Government funded training for the unemployed	
Impact Evaluation Score: 4 (see end of summary)	**Monetisation Score: n/a** (see end of summary)

Time period covered by policy: On-going	**Time period covered by evaluation:** Data relating to new unemployed benefit claimants between 2006-08 and 2010-12
Contractor undertaking evaluation: David Bibby, Augusto Cerqua, Dave Thomson and Professor Peter Urwin (Centre for Employment Research, University of Westminster)	**Peer reviewers:** Professor Anna Vignoles (University of Cambridge) Steven McIntosh (Reader in Economics, University of Sheffield)
Type of evaluation: Outcome evaluation - using an administrative database containing the records of all unemployed benefit claimants from DWP matched with any publically funded FE learning records from BIS and their employment and pay records from HMRC to enable observation and comparison of labour market activity before and after learning.	

Description of policy/programme and rationale for intervention:

Government funded training for the unemployed seeks to support unemployed individuals who need retraining to help them get back into the labour market.

Summary of key evaluation findings:

- Evidence found of positive and statistically significant employment and sustained employment premiums for all FE qualification categories when taken by the unemployed
- In all qualification categories some evidence that achievers have a statistically significantly lower probability of being on active benefits than non-achievers.
- Estimates of value added for 18-24 year olds, based on a comparison of

outcomes between FE achievers and a matched control group who do not undertake FE learning confirm the findings of good return to FE learning.

Summary of cost-benefit/cost-effectiveness analysis (if applicable):

N/A

Policy response to the evaluation:

For the first time, the research provides evidence on the returns to FE for the unemployed including for specific levels and types of training, and allows us to understand the impacts for the long-term unemployed as well as short-term unemployed. Key messages and policy implications are being feed into in policy development for the 2015 Spending Review.

Evaluation methodology

Description of methodology: Evaluation uses matched administrative data on wage, employment and FE learning following a claim start date to estimate the value added for different forms of FE learning, by comparing returns over the 60 months from claim start date, between (i) 'ILR achievers' (ii) 'ILR non-achievers' and (iii) those with 'no-ILR record'. Individuals are matched according to their unemployment histories and returns are differentiated according to the duration of their unemployment spell.
Does the evaluation review the published policy objectives? N/A
At what level are the main intended outputs and/or outcomes expected to occur? (What is the unit of analysis? For example: universities, businesses, individuals or nationally) Individuals
Has sufficient time lapsed for the initial/full benefits to be estimated? Yes – labour market outcomes are observed for 1-3 years following the intervention (please note outcomes are observed up to 4 years from learning start date)

Peer review

Comments on the appropriateness of data and outcomes: **Anna Vignoles:** *The data uses high quality data on unemployment linked to census data on training and qualifications acquired. The weakness of the data, as is the case with any administrative data, is that the covariates that can be included in any model of employment are necessarily quite limited since the data are not rich. The data do not, for example, have good information on individuals' cognitive skill prior to the unemployment spell. This potentially threatens internal validity but the authors have*

used an appropriate range of econometric techniques to try to overcome this. Overall, the data set has clear advantages and provides a high quality outcome measure.

Steven McIntosh:

The aim of this study is to determine the impact of various aspects of FE learning and training, on labour market outcomes, namely likelihood of employment and likelihood of benefit receipt. The focus of the study is, more specifically on the impact of low level learning, up to level 2, for learners who are initially unemployed. This is a wholly valid question to ask. Unemployment and benefit receipt remain major policy issues for a whole range of reasons, for example social mobility, poverty, efficient functioning of the economy, public debt etc. Youth unemployment has merited its own attention, given its higher level than adult unemployment, and the long-lasting scarring effects of a lack of work early in one's working life. The analysis is therefore rightly divided into those aged 18-24 and those aged 25+ throughout the report. The question of the impact of vocational and training is also an appropriate one, given that the acquisition of new skills is an oft-advocated route out of unemployment, particularly amongst the young.

Furthermore, one of the difficulties of investigating the labour market effects of low level vocational qualifications has been finding a suitable counterfactual amongst the general population. Restricting the analysis to those out of work is a first step in making treatment and control groups more similar (see discussion in next section).

The study uses matched data from the Individualised Learner Record (ILR) of learners in FE, and the Work and Pensions Longitudinal Study (WPLS). At various points, data from the New Deal (ND) programme are also included. Thus, all of the analysis is based on administrative data, following several other recent studies described in the report that use the same data. This contrasts with earlier work in the area which was traditionally survey-based, for example using data from the Labour Force Survey. The use of administrative data has a number of advantages, in particular extremely large numbers of observations typically covering the population rather than a surveyed sample. In addition, administrative data typically include much more detailed information on the type of learning, allowing for more disaggregated analysis, for example by type of qualification, level and sometimes subject.

Comments on internal validity:

Anna Vignoles:

The authors compare employment of achievers versus non-achievers to identify the impact of a particular qualification level. This approach is, as the authors fully

acknowledge, potentially problematic since those who fail to achieve a qualification may by definition be different in unobservable ways to those who complete one. For robustness therefore, the authors also compare achievers with those who are not participating in FE at all and the results, if anything, suggest an even greater employment effect when they use this latter approach. The strength of the matching and in particular their controls for prior spells and duration of unemployment support the robustness of their results. There are however still remaining possible sources of biases.

Overall, the report has done as much as is feasible given data limitations to overcome internal validity threats.

Steven McIntosh:

This report follows several other recent studies that use the matched ILR-WPLS dataset to study the returns to low level vocational qualifications. It differs from earlier studies in that it considers a specific group of individuals, namely those identified as unemployed and beginning a benefit spell within the 2006-8 period (with further analysis on a later cohort). The individuals who acquire a low level vocational qualification are compared to two separate control groups, specifically individuals who similarly begin in a benefit spell within the same period and who (i) begin the same learning aim but do not complete, and (ii) who are not observed in the ILR over the relevant period beginning any learning aim.

Considering the first control group of non-achievers first, similar control groups have been used by a number of studies, referenced in the report, using the same matched administrative data set. The big advantage of using the non-achievers as the control group is that they at least registered for the same learning aim as the treatment group. Thus they have the same motivation, and need, to acquire the skills provided by low level vocational qualifications, and also have the same unobservable characteristics that might affect labour market outcomes, and also contribute to the decision to undertake such qualifications. In short, the control group 'look like the sort of individuals who undertake low-level vocational qualifications', and so can be argued to be a more appropriate control group than any found in the general, full population. For example, the group of individuals in the full population with no qualifications are not a good source of comparison groups, because of the likely variability in their skills, abilities, motivations, ages and experiences.

The argument put forward against the use of non-achievers as a control group is that they may be fundamentally different from the treatment group due to the very fact that they have failed to achieve the learning aim. Thus, so the argument goes, the employment outcomes of the achievers are better, because they were more able or motivated enough to achieve their learning aim in the first place.

There are various ways in which the report attempts to counter this argument. First, the treatment and control groups come from the same, relatively narrowly-defined, population, namely individuals who are observed to have begun a spell of unemployment within the sample period. Furthermore, the short-term and long-term unemployed, who might have very different skills, experiences and needs, are separated and matched across treatment and controls independently of each other.

The individuals being studied therefore do not belong to a widely-defined population with very different characteristics.

Second, and most importantly, the treatment group and control group are exactly matched (using Coarsened Exact Matching, CEM) according to key characteristics. One of the criticisms of the use of administrative data in the past has been that there are too few individual characteristics with which to match treatment and control groups. I do not think that this criticism implies in this case though, as there is a quite impressive list of characteristics on which to match. Most importantly, the authors have information on labour market experience (specifically, whether the individual was in employment in month (t-1); in employment month (t-2); in employment month(t-6); and number of months in employment for various periods between months (t-7) and t-60)). Thus the analysis compares treated achievers to those non-achievers who have performed similarly in the labour market prior to treatment. Furthermore, after the matched sample was created, a standard regression equation was estimated, controlling additionally for whether an unemployed individual has children; ethnicity; whether the individual is a previous offender; age; lone parent status; and number of prior LMS opportunities and prior ILR aims started.

The success of the matching process is shown in various figures throughout the report that show the differences between the treatment and control groups for the various outcome variables, in the five years before the treatment group were treated. The figures show, in every case considered, that the outcome variables were very similar, and insignificantly different from each other, between treatment and control groups before the treatment group were treated. Thus, the individuals in the two groups had the same employment and benefit receipt likelihoods, before the achievers started their learning.

The final way in which the report counters the argument that achievers and non-achievers are fundamentally different in unobservable ways is that it also uses the second control group of non-ILR learners, i.e. individuals starting a benefit claim in the relevant period, but not appearing in the ILR as having started a learning aim. I consider the use of this additional control group as a good step forward in this sequence of reports using the matched administrative data sets. Recall that the aim of the use of both control groups is to estimate the same thing, i.e. the counterfactual of what would have happened to the treatment group of learning achievers, if they had not successfully completed their course. As such, if both control groups provide accurate, unbiased estimates, then they should give the same estimate of the treatment effect. If they do not, then we would not know which, if indeed either, was the correct estimate. The fact that the estimated treatment effects are very similar across control groups, for the 18-24 year old age group at least, lends support to the claim that they are estimating the true effect of the learning. For this not to be the case, the same confounding effect, correlated with treatment but one-off in the sense that it did not exist before treatment, would need to affect the treatment group relative to both control groups equally.

Comments on external validity:

Anna Vignoles:

The estimate of the employment effect is for both the short term and the long term unemployed. Clearly these are select groups. This limits the external validity of the study in the sense that it cannot be used to judge the employment benefits of NVQ2 for those who do not experience unemployment. This is not really a weakness however since understanding the impact of qualifications on the employment prospects of the short and long term unemployed is a crucially important policy question.

Steven McIntosh:

The analysis in this report is based on the population of a particular group of individuals, namely those individuals identified with a first or only (job-seeking) benefit claim (i.e. unemployed) in the period between April 2006 and April 2008. A second, later cohort, similarly defined for the period between August 2011 and July 2012 is also considered. There is therefore no question about the representativeness of this data set, for the population being considered, since the data set contains that full population. The results, on the first cohort for example, tell us the effect of learning on labour market outcomes of those who are observed as unemployed between 2006 and 2008. Can the results be generalised outside of this population? In terms of time period, the analysis using the second, 2011-12 cohort, finds very similar results, suggesting that the findings are generalisable across cohorts, and that the main analysis of the report from 2006-8 can still be applied to the present day. In terms of whether the results can be applied to others outside of the population, such a question would be missing the point of the study somewhat. The study is not saying that low level qualifications will benefit all individuals in the full population. Many individuals would receive no benefit, or even suffer negative consequences, from acquiring such qualifications (the earlier studies based on survey data covering the full population have shown such results). However, amongst the 'sort of people amongst the unemployed who acquire low level vocational qualifications,' then they can have beneficial effects on labour market outcomes, compared to not acquiring such qualifications'.

Comments on the quality of inferences and establishing causation:

Anna Vignoles:

The authors have done a good job caveating their interpretation of the results notwithstanding the comments above. Despite some threats to internal and external validity, it is a robust piece of work that improves our understanding of the issues.

Steven McIntosh:

As discussed in the section on Internal Validity above, the authors have taken various steps to reduce the possibility that confounding effects are actually causing the difference in outcomes between treatment and control groups. To what extent then, can we interpret the results as causal effects? I think that the authors have gone a long way to establishing causality. They compare groups of achieving

learners to both non-achieving learners and non-learners from a relatively narrow population, namely those entering either short-term or long-term unemployment (analysed separately). The use of the pre-learning data is important here, and the authors successfully show that these groups had the same labour market experiences, in terms of employment and benefit receipt likelihoods, prior to learning. The analysis then also controls for a range of other socio-economic and prior learning characteristics. I consider this list of variables matched on and controlled for impressive in terms of its number, and also because these are exactly the variables that we would want to be matching on and controlling for. The longer this list of variables used, the smaller the probability that there remains another variable that differs between treatment and control groups that actually explains the difference in their outcomes. As an example, in an ideal world, it would be helpful if data from the National Pupil Database (NPD) could be matched in to the merged ILR-WPLS data set, so that childhood educational achievements could also be held constant between treatment and control groups. On the other hand, for them to be causing the observed results, given that they are not being controlled for, they would need to have no impact on relative outcomes between treatment and control groups before learning, with an effect only emerging to differentiate outcomes after the former engage in learning. While it is not impossible for such a situation to occur (for example through synergies between childhood and FE learning) the causal interpretation remains in my mind a more likely explanation.

Is there anything else the authors can do to justify a causal interpretation? As I have argued above, they have already done a lot. One possibility is to look at the reasons for non-achievement of FE learning aims, either in the current data set or in previous literature. I agree with the authors' careful naming of this group as 'non-achievers' rather than 'failures'. Failure of the end assessment is only one reason for the non-achievement of learning aims, and the more that non-achievement can be shown to be due to random events outside the learners' control, for example, characteristics of the course enrolled on, rather than due to endogenous characteristics of the learners themselves, then the even stronger the weight that can be placed on the causal interpretation.

Other comments:

Steven McIntosh:

In terms of presentation, I thought the design, results and implications of the study were all clearly explained. What the authors were doing, and why they did it, were made clear to the reader throughout. Obviously there are a lot of different specifications and results presented in the report. I therefore found the summary tables of results in the Executive Summary extremely useful, allowing the reader to see all the results at a glance, and to easily compare across them.

Cost-effectiveness and cost-benefit summary

Justification for monetisation score: N/A
Sensitivity analysis/key assumptions: N/A

Direct costs to Exchequer of programme:

£m	Total	Year 0	Year 1	Year 2
Total				

Economic costs and benefits of programme:

Price base year	2013/14	**Present value base year**	2013/14	**Discount rate**	3.5%

	Costs (£m)			Benefits (£m)			NPV (£m)	Net BCR[1]
	Transition (constant price)	Average annual	Total (PV)	Transition (constant price)	Average annual	Total (PV)		
Low								
Best estimate								
High								

Description and size of key monetised costs: N/A
Other key non-monetised costs: N/A
Description and size of key monetised benefits: N/A
Other key non-monetised benefits: N/A
Robustness of monetised costs and benefits: N/A

Peer Review

Evaluation peer review comments on comprehensiveness, clarity, robustness and best practice of cost benefit/cost effectiveness analysis: N/A

[1] PV of net benefits / PV of net costs

Note on Impact Evaluation and Monetisation Scores

Impact Evaluation Score

Impact scale follows the guidance on 'Quality on Impact Evaluation[2]', published as supplementary guidance to the Magenta Book. The scale is based largely on the Maryland Scientific Method Scale used by academics and researchers to assess the strength of an evaluation approach. The higher the score potentially the more capable the evaluations are to demonstrate that the outcome observed is due to or caused by the intervention.

- Score 5: Random allocation of treatment and control group, or a robust counterfactual using a quasi-experimental approach. There is a treatment and a comparison group and actual before and after data in both groups. For example: a strong difference-in-difference design, regression discontinuity design or matched treatment and control group.

- Score 4: Quasi-experimental approach where the counterfactual has some weaknesses, but it is as good as can be, given the policy design or data availability issues. There is a treatment and a comparison group, and actual before and after data in both groups. For example: a difference-in-difference design, regression discontinuity design or matched treatment and control group.

- Score 3: Predicted (modelled) versus actual outcomes for the treatment group only are compared, predictions are based on actual baseline data.

- Score 2: Actual (i.e. not self-assessed or self-reported) before and after data for the treatment group only are compared. (Higher levels on this scale also require actual data not based on self-reporting.)

- Score 1: No baseline data (or only self-assessed/self-reported data).

Monetisation Score

The higher the score the more information the evaluation contains in terms of analysing the cost of the intervention and the additional benefits to the economy.

- Score 5: Input, output, outcome data additional Benefit Cost Ratio (BCR), NPV set aside some other not monetised impact measures, fuller cost benefit analysis or cost effectiveness analysis that compares the costs of alternative

[2] Quality in policy impact evaluation, HMT, Dec 2012

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/190984/Magenta_Book_quality_in_policy_impact_evaluation__QPIE_.pdf

ways of producing the same or similar outputs

- Score 4: Input, output, outcome data, calculation of additional Benefit Cost Ratio, Net Present Value
- Score 3: Input, output, outcome data calculation of Gross BCR not additional or not clear if additional
- Score 2: Gross BCR not available, as either input or output data are not available
- Score 1: No monetisation at all