THE CAMPBELL COLLABORATION

The effect of benefit exhaustion on unemployment.

Trine Filges, Lars Pico Geerdsen, Anne-Sofie Due Knudsen, Anne-Marie Klint Jørgensen and Krystyna Kowalski

Submitted to the Coordinating Group of:

- ☐ Crime and Justice
- ☐ Education
  - ☐ Disability
- ☐ International Development
- ☒ Social Welfare
- ☐ Other:

Plans to co-register:

- ☐ No
- ☐ Yes ☐ Cochrane ☐ Other
- ☐ Maybe

Date Submitted: 26 Nov 2010

Date 1st Revision Submitted: June 2011

Date 2nd Revision Submitted: 08 September 2011

Approval Date:14 October 2011

Publication Date:

Table of contents

1 Background

1.1 DESCRIPTION OF THE CONDITION

In 1970 the unemployment rate in the US was 5% while the unemployment rate in the European Union was 3% (Solow, 2000). Since the first oil crisis in 1973, the unemployment rate in Europe and the US has diverged. While it remained relatively steady in the US there was an upward trend in Europe. By the end of the century the unemployment rates in most European countries did not seem to go back to the low levels that were commonplace thirty years ago (the average unemployment rate in the European Union was around 10%).[1] This steady contrast has posed the inevitable question: what explains the difference between the levels of unemployment in Europe and the US? Many explanations have been put forward to explain this difference but no single factor has been traced as the explanation. In labour market research, the conventional understanding is that the difference rests on differences in labour market "institutions." The variables considered are, among others, the unemployment benefit system, trade union power, taxes, employment protection, barriers on labor mobility, and wage inflexibility (Layard et al., 2005; Nickell et al., 2005). Among these variables, the benefit system is shown to be one of the key factors (Layard et al., 2005)[2]. The main aspect of the benefit system that influences unemployment is the generosity of the system either in amount or in duration. In the US, replacement rates[3] are low and duration is short compared to most European countries. The natural consequence is that higher levels of active searches and a greater willingness to accept inferior jobs by unemployed workers are seen in the US than in Europe.

From a societal point of view, the optimal benefit system is determined as a trade-off between protection and distortion. Benefit programs protect individuals against loss of income and provide unemployed individuals the possibility of finding a better match between their qualifications and job vacancies. In fact, this positive aspect of inducing risk-averse workers to achieve better job matches has been shown to

[1] By the end of the previous decade however the financial crisis turned things upside down. The US unemployment rate was 10% in 2010 whereas the unemployment rates in major European countries as Italy, UK and Germany was lower than in the US. The average unemployment rate in the European Union was however at the same level as in the US.

[2] Other key factors are the co-ordination level of the wage bargaining and employment adjustment costs.

[3] The replacement rate is the ratio of the unemployment benefit and previous earnings.

increase economic efficiency (Acemoglu & Shimer, 1999; Marimon & Zilibotti, 1999). However, the same benefit can also distort incentives through job searches that are long and unproductive. Therefore, unemployment benefits should aim for a balance between protection and distortion (Feldstein, 2005; Mortensen, 1987).

In order to reduce the high unemployment level, European policymakers may wish to reduce the generosity of the unemployment system. While it may be politically intractable to lower the replacement rate, (indeed, examples of reductions of benefit rates and amounts are rare), the length of the unemployment benefit eligibility period is often used as a political instrument to create work incentives for the unemployed. For example, the benefit period was altered in Spain in 1992, in Slovenia in 1998, in Norway in 1997, in the UK in 1996, in Denmark in 1996, 1998 and 1999, and, more recently, in the Czech Republic in 2004, in Hungary and Portugal in 2006, and in Denmark in 2010.

This review will focus on the effect of exhaustion of unemployment benefits and look at the unemployed workers' exit rate into employment *prior* to exhaustion of unemployment benefits or shortly thereafter. The effect occurring *prior* to benefit exhaustion or shortly thereafter, which we will denote the "incentive effect," is relevant because several studies, empirical as well as theoretical, suggest that the prospect of exhaustion of benefits results in a significantly increased incentive for finding work (Card, Chetty, & Weber, 2007; Caliendo, Tatsiramos, & Uhlendoff, 2009; Feldstein, 2005; Katz & Meyer, 1990; Meyer, 1990; Mortensen, 1987). Hence, shortening the benefit eligibility period may reduce the share of long and unproductive job searches and thereby decrease the overall unemployment level.

1.2 DESCRIPTION OF THE INTERVENTION

The intervention is the exhaustion of any kind of unemployment benefit with a known expiration date. The review will focus on the incentive effect, i.e., the exit rate out of unemployment into employment *prior* to exhaustion of unemployment benefits or shortly thereafter, which can be attributed solely to the prospect of benefits exhaustion.

The benefits may be unemployment insurance (UI) benefits or they may be unemployment assistance (UA)/social assistance (SA) benefits as long as they have a known expiration date.

In the majority of OECD countries, the unemployment insurance (UI) benefit has a time-limit. In fact, only Belgium has an unlimited UI period. In other countries, the maximum duration varies between 6 months (as for example in the UK and the US) and 36 months (in Iceland). In Denmark the maximum duration was lowered from 48 months to 24 months in 2010.

In most OECD countries, a secondary benefit is available for those who have exhausted regular UI benefits. This is known as social assistance (SA) benefits. Unlike UI benefits, SA benefits are generally means-tested, pay a lower level of benefit, and are indefinite. We know of only one example of a SA benefit with a time limit. This is the Temporary Assistance to Needy Families (TANF) available in the US. The federal government requires states to impose between two or five year limits (see Gustafson & Levine, 1997). In a minority of OECD countries, unemployment assistance (UA) benefits are paid after exhaustion of UI benefits. Like SA benefits they are generally means-tested, pay a lower level of benefits, and, excepting Hungary, Portugal, and Sweden, they are generally indefinite.

The main control or comparison condition is benefit exhaustion that is not immediate.

1.3 HOW THE INTERVENTION MIGHT WORK

Search theory offers an explanation for why we expect to find an effect of this intervention. According to search theory, one can derive a relationship between the job-finding rate and the time to benefit exhaustion when the maximum benefit duration is fixed and predictable (Mortensen, 1977). This relationship is driven by adjustments in search effort and reservation wages. The reservation wage is the minimum wage at which the unemployed are willing to accept a job. The benefit exhaustion gives the unemployed individual a strong incentive to gain employment to avoid the drop in income after the exhaustion date. How strong the incentive is depends on the magnitude of the income drop. If no secondary benefit is available for those who have exhausted their current benefit, the incentive to gain employment will be stronger. As the unemployed worker approaches benefit expiration, the search intensity goes up and the reservation wage goes down, thus increasing the job finding rate. If an increased job finding rate is mainly driven by lowering the reservation wage, a lower job match quality is to be expected, for example, in the form of lower wages and/or lower re-employment duration.

A number of factors may have an impact on the magnitude of the expected increase in the job finding rate when approaching benefit exhaustion. In general, the overall labour market conditions, i.e. the vacancy rate and, in particular, the unemployment rate, have an impact on the availability of and competition for jobs.

The maximum benefit duration is also expected to have an impact on the size of the exhaustion effect. The higher the initial benefit eligibility period, more sorting may be expected to have taken place and, hence, a smaller benefit exhaustion effect is expected. Sorting refers to a dynamic selection mechanism based on a relationship between individual heterogeneity (i.e. heterogeneity in the individual characteristics of the unemployed) and the hazard of leaving unemployment. Heterogeneity is related to job performance; those perceived to be most productive and more desirable to employers are hired first (Salant, 1977 and Jackman & Layard, 1991).

Several studies find sorting effects. For example Lancaster, 1979 and Narendranathan & Stewart, 1993 and more recently the analysis in Kalwij (2010) identifies significant sorting effects. They show that both observed (to the researcher) heterogeneity (e.g. age and education) and unobserved (to the researcher) heterogeneity (e.g. motivation and 'drive') are important determinants of the unemployment hazard.

The extent to which those left unemployed by the end of the benefit eligibility period are considered unproductive and not desirable to employers has an effect on their unemployment hazard and therefore has an impact on the exhaustion effect (i.e. it may be impossible to find an employer willing to hire the unemployed no matter how high the search intensity is and how low the reservation wage is).

Whether compulsory participation in active labor market programmes is part of the unemployment system may result in additional sorting. The compulsory aspect may provide an incentive for unemployed individuals to look for and return to work prior to programme participation (Geerdsen et al., 2011). Further, participation in active labor market programmes may improve some of the participants' qualifications helping them to find a job. Hence, those left unemployed by the end of the benefit eligibility period may be considered even more unproductive if participation in active labour market programmes did not improve their qualifications and lead to a job. Alternatively, active labour market programmes may have negative stigmatization and signaling effects to employers. Programmes associated with participants having poor employment prospect may carry a stigma. Because of asymmetric information, employers do not know the productivity of new workers, some of whom they might hire from the pool of the unemployed. Prospective employers might then perceive participants in such programmes as low productivity workers or workers with tenuous labor market attachment (Kluve et al. 1999; Kluve et al., 2007).

Finally, the type of unemployment benefit may have an impact on the job finding rate close to exhaustion. As mentioned above, some countries employ two systems to provide benefits to unemployed individuals: an unemployment insurance system for individuals who typically have a strong labour market attachment (UI benefits) and a social welfare system for individuals who often have other problems in addition to un-employment (SA or UA benefits). The effect size in social welfare systems offering unemployment benefits with a known expiration date is expected to be less than the effect size in unemployment insurance systems with a known expiration date.

1.4 WHY IT IS IMPORTANT TO DO THIS REVIEW

There are many empirical papers on the effect of benefit exhaustion on unemployed individuals (Caliendo, Tatsiramos and Uhlendoff 2009; Card, Chetty and Weber 2007; Katz and Meyer 1990; Lalive et al 2006; Meyer 1990), but the empirical

research has not been summarized in a systematic review to obtain a clearer picture of the available evidence on the employment effect of benefit exhaustion. One paper provides a review of the recent literature on how incentives in unemployment insurance can be improved (Fredriksson and Holmlund 2006). However, it is not a systematic review and, furthermore, the authors do not make the important distinguistion between exits to employment and exits to other destinations such as secondary unemployment benefits. Distinguishing between destinations is vital. As shown in Card, Chetty and Weber (2007), the exit rate from registered unemployment increases over 10 times more than the rate of re-employment at the expiration of benefits. The difference between the two measures arises because many individuals leave the unemployment register immediately after their benefits expire without returning to work.

There is a great deal of political interest in optimizing the unemployment benefit system so it balances the protection and distortion dimensions. The political interest is to reduce the unemployment level, to prevent exploitation of the unemployment benefit system and at the same time protect the unemployed individuals with real difficulties in finding a job. It is therefore of great importance to find out what effect unemployment benefit exhaustion has on employment probabilities.

2 Objective of the Review

The primary objective of this systematic review is to study the impact of exhaustion of unemployment benefits. The primary outcome is unemployed individuals' exit rate out of unemployment and into employment prior to benefit exhaustion or shortly thereafter.

3 Methods

3.1 CRITERIA FOR CONSIDERING STUDIES FOR THE REVIEW

3.1.1 Types of Studies

The study designs included in the review are:

- Controlled trials (all parts of the study are prospective, i.e. identification of participants, assessment of baseline, allocation to intervention, assessment of outcomes and generation of hypotheses, see Higgins & Green, 2008):
    - RCT - randomized controlled trial
    - QRCT - quasi-randomized controlled trial (i.e. participants are allocated by means such as alternate allocation, person's birth date, the date of the week or month, case number or alphabetical order)
    - NRCT - non-randomized controlled trial (i.e. participants are allocated by other actions controlled by the researcher)
- Non-randomized studies (includes truly observational studies where the use of an intervention has occurred in the course of usual treatment decisions or peoples' choices)
    - NRS - the allocation is not controlled by the researcher and there is a comparison of two or more groups of participants. Participants are allocated by means such as time differences, location differences, decision makers, policy rules or participant preferences.

We will include study designs that use a well-defined control group. Studies that utilize qualitative approaches will not be included in the review due to the absence of adequate control group conditions.

Notably due to the nature of the field few randomised controlled trials are performed regarding the effectiveness of such social and labour market policies. Therefore, studies of the effect of benefit exhaustion typically are estimated on observational data, often collected from administrative registers or by questionnaires. Studies that use different data sources for treatment and control groups will not be included in this review.

3.1.2 Types of Participants

The participants will be unemployed individuals who receive some sort of time limited benefit during their unemployment spell. We will include participants receiving all types of unemployment benefits with a known exhaustion date. The only restriction is that the benefits must be related to being unemployed. We will therefore exclude individuals receiving other types of benefits not related to being unemployed. We will not restrict our attention to certain types of participants, since the main focus of this review is on the incentive effect to find a job when benefits exhaust. Therefore, we will include all unemployed participants regardless of age, gender, etc. who receive some sort of time limited benefit during their unemployment spell.

3.1.3 Types of Interventions

The intervention of interest is the exhaustion of any kind of unemployment benefits with a known expiration date. The review will focus on the incentive effect, i.e. the exit rate out of unemployment into employment *prior* to exhaustion of unemployment benefits or shortly thereafter. The benefits may be unemployment insurance (UI) benefits or they may be unemployment assistance (UA)/social assistance (SA). The only requirement is that the benefit must have a known expiration date. The UI benefit usually has a known time-limit whereas UA and SA usually are indefinite. Unemployment benefits with an indefinite time limit or non-financial benefits will be excluded from this review. The comparison will be benefit expiration that is not immediate.

3.1.4 Types of Outcomes

The objective of the review is to determine whether the prospect of unemployment benefit exhaustion motivates unemployed individuals to find a job. Distinguishing between destinations is therefore vital. The primary outcome is exits to employment. Studies only looking at exits to other destinations such as other types of social benefits or non-employment will not be included in this review or studies who do not distinguish between destinations will be excluded from this review.

In addition to the primary outcome measure, we will consider secondary outcomes in terms of the impact the exhaustion of benefit has on the duration of employment and on income. This is to obtain a clearer picture of the effect the prospect of benefit exhaustion has on the quality of the job. If the duration of re-employment or the wage is low, this could indicate that the exhaustion of benefits forces unemployed individuals to find jobs that do not match their qualifications and therefore they may return to unemployment quickly.

Primary outcomes

a) Exit rate from unemployment to employment

Secondary outcomes

a) Duration of re-employment

b) Re-employment wage

3.2 SEARCH METHODS FOR IDENTIFICATION OF STUDIES

3.2.1 Electronic Searches

Relevant studies will be identified through electronic searches of bibliographic databases, government policy databanks and internet search engines. No language or date restrictions will be applied to the searches.

Business Source Elite
EconLit
PsycInfo
SocIndex
Science Citation Index
Social Science Citation Index
The Cochrane Library (Cochrane reviews, other reviews, clinical trials)
International Bibliography of the Social Sciences
IDEAS/Economist Online/Social Care Online [4]
Dissertation Abstracts International
Theses Canada

3.2.2 Search Terms

An example of the search strategy for Business Source Elite is listed below. The strategy will be modified for the different databases. We will report full details of the modifications in the completed review. As this review also includes non-randomised study designs (i.e. quasi-experimental designs) trial filters are not relevant.

1. DE "Social Security"
2. DE "WELFARE recipients"
3. welfare w1 payment*
4. welfare w1 recipient*
5. welfare w1 support*
6. economic w1 support*
7. public w1 assistance*
8. welfare w1 payment*
9. public w1 support*
10. financial w1 support*
11. welfare w1 service*
12. direct* w1 payment*
13. general w1 assistance
14. Social w1 Support
15. cash w1 assistance

[4] The search strategy will have to be modified to a great extent to searching these databases. The search interfaces does not allow complex searching. Even though these databases contain similar references, we intend to search both trying to perform as thorough a search as possible.

16. income w1 assistance
17. benefit*
18. social w1 assistance*
19. social w1 securit*
20. social w1 welfare
21. social w1 allowance*
22. insurance w1 benefit*
23. social w1 benefit*
24. welfare w1 benefit*
25. TANF
26. Insurance*
27. 1-26/or
28. DE "EMPLOYABILITY"
29. Employ*
30. Job*
31. work*
32. un-employ* or unemploy*
33. re-employ* or reemploy*
34. 28-33/or
35. effect*
36. threat*
37. incentive*
38. disincentive*
39. impact*
40. motivat*
41. 35-40/or
42. Expir*
43. Lapse*
44. Terminat*
45. Duration*
46. Generosit*
47. Change*
48. Entitl*
49. Length
50. Extend*
51. Extension*
52. Exhaust*
53. exit*
54. 42-53/or
55. 27 and 34 and 41 and 54

3.2.3 Hand Searching in the Following Journals

Reference lists of included studies and reference lists of relevant reviews will be searched. “The Journal of Labor Economics” and “Labour Economics” will be hand searched for the year 2010 and the available issues of 2011.

3.2.4 Other Sources

Citation searching in Web of Science of included studies will also be considered.

3.2.5 Grey Literature

Google will be used to search the web to identify potential unpublished studies. Advance search options will be used to refine the grey search strategy. OpenSIGLE will be used to search for European grey literature (http://opensigle.inist.fr/).

The private independent research institutes and economic networks:

IZA – Institute of the Study of Labor (www.iza.org)
CEPR – Centre for Economic Policy Research (www.cepr.org)
NBER – National Bureau of Economic Research (www.nber.org)
MDRC – the Manpower Demonstration Research Corporation – (www.mdrc.org)
CESifo – the cooperation between CES (Center for Economic Studies) and IFO (Institute for Economic Research) – (www.cesifo-group.de/portal/page/portal/ifoHome) are all covered via IDEAS.

In addition we will search:
Danish Economic Councils (www.dors.dk)
OECD - the Organisation for Economic Co-operation and Development (www.oecd.org)
IMF - The International Monetary Fund (www.imf.org)
AIECE - Association of European Conjuncture Institutes (www.aiece.org)
ESRC - Economic Social Research Council (www.esrc.ac.uk)
Copenhagen Economics (www.copenhageneconomics.com)

SSRN – Social Science Research Network (www.ssrn.com) will also be searched to uncover potential preprint discussion papers.

Unpublished theses and dissertations will be searched through the databases: Theses and dissertations and Theses Canada.

Copies of relevant documents will be made recording the exact URL and date of access.

3.2.6 Personal Contacts

Personal contacts with national and international researchers will be made to identify unpublished reports and on-going studies.

3.3 METHODS USED IN PRIMARY STUDIES

Studies of the effect of benefit exhaustion are typically estimated on observational data, often collected from administrative registers or by questionnaires. In these studies, individuals reaction to the exhaustion of benefits is analysed using either panel data or duration data. By using data which describes individuals over time, it is possible to see how people move between the different states in the labour market.

The central problem in these studies is the identification of the incentive effect. Often the variable describing time to benefits exhaustion is a function of variables which all have a direct effect on individuals' duration of unemployment. But, identification necessitates that at least one of the variables be omitted from the modelling of individuals' unemployment duration (the exclusion restriction)[5]. Examples of exclusion restrictions used in the literature are differences in benefits entitlement over regions or between individuals due to previous unemployment, see Ham & Rea, 1987 and Meyer, 1990. In order to use this variation to disentangle the incentive effect from other time varying effects, one has to assume that this variation does not on its own have an effect on individuals' unemployment duration. More recently, in a number of studies identification is driven by legislative changes of the maximum entitlement period, see e.g. Boone & van Ours, 2009 and Lalive, van Ours & Zweimüller, 2006. This makes it possible to compare individuals' labour market behaviour just before and after the change was implemented.

The incentive effect in these studies is given by the difference in hazard rates *prior* to benefit exhaustion or shortly thereafter between persons who approach exhaustion and persons who do not approach exhaustion. Some of the studies estimate the incentive effect using indicator variables for the number of months or weeks until exhaustion (Boone & van Ours, 2009 and Lalive, van Ours & Zweimüller, 2006, whereas Meyer, 1990, uses a spline function and Ham & Rea, 1987, uses a log polynomial function describing the same time period.

3.4 DATA COLLECTION AND ANALYSIS

3.4.1 Selection of Studies

Two reviewers (ADK, TF) will independently read titles and available abstracts of reports and articles identified in the search to exclude reports that are clearly irrelevant. Citations considered relevant by at least one reviewer will be retrieved in full texts versions. If there is not enough information in the title and abstract to judge relevance, the full text will be retrieved. At least two reviewers (TF, ADK) will read the full text versions to ascertain eligibility based on the selection criteria. In the first screening level (on the basis of title and abstract) a citation will only move on to the second screening level if the answer is a yes or uncertain for the following

[5] Alternatively identification can be achieved by assuming that the duration dependence follows a specific functional form. For more information see Geerdsen, 2002.

criteria: (1) Are the participants' unemployed individuals receiving some kind of benefit during their unemployment? (2) Does the study focus on time limits in the unemployment benefit eligibility period? (3) Is the report/article a quantitative evaluation study?

In the second level (on the basis of full text) eligibility inclusion criteria are extended to the following: (4) Does the study estimate an effect, using a control group or using an estimated counterfactual? (5) Does the study examine exits to employment?

The inclusion criteria questions for level 1 and 2 will be piloted and adjusted if required (see appendix 1). Primary investigators will be contacted to clarify study eligibility if necessary. In the event of disagreements, a third reviewer and content specialist (LP) will be consulted and consensus will be sought. Exclusion reasons for studies that otherwise might be expected to be eligible will be documented and presented in an appendix. The overall search and screening process will be illustrated in a flow-diagram. None of the review authors will be blind to the authors, institutions, or the journals of publication of the articles.

3.4.2 Data Extraction and Management

At least two review authors (TF, ADK) will independently code and extract data from the included studies. A data extraction sheet will be piloted on several studies and revised as necessary (see appendix 2). Extracted data will be stored electronically. Disagreements will be resolved by consulting an independent reviewer with extensive content and methods expertise (LP). Analysis will be conducted in RevMan5, SAS, and Stata. Data and information will be extracted on: characteristics of participants, intervention characteristics and control conditions, research design, sample size, risk of bias and potential confounding factors, censoring, outcomes and results.

3.4.3 Assessment of Risk of Bias in Included Studies

We will assess the methodological quality of studies using a risk of bias model developed by Prof. Barnaby Reeves in association with the Cochrane Non-Randomised Studies Methods Group.[6] This model, an extension of the Cochrane Collaboration's risk of bias tool, covers both risk of bias in RCTs, but also risk of bias in non-randomised studies that have a well-defined control group.

The point of departure for the risk of bias model is the Cochrane Handbook for Systematic Reviews of Interventions (Higgins & Green, 2008). The existing Cochrane risk of bias tool needs elaboration when assessing non-randomised studies because, for non-randomised studies, particular attention must be paid to selection bias/risk of confounding. It is also important to try to discriminate between non-

[6] This risk of bias model was introduced by Prof. Reeves at a workshop on risk of bias in non-randomised studies at SFI Campbell, February 2011. The model is a further development of work carried out in the Cochrane Non-Randomised Studies Method Group (NRSMG).

randomised studies with varying risk of bias, so the model requires assessment on a 5-point scale for some items.

3.4.3.1 Risk of Bias Judgement Items

The risk of bias model is based on 9 items (see appendix 7.3). Some items are judged by High/Low/Uncertain and some by a 5-point scale. Using the 5-point scale, 1 corresponds to No/Low risk of bias and 5 correspond to Yes/High risk of bias. Five corresponds to a risk of bias so high that the findings will not be considered in the data synthesis (because they are more likely to mislead than inform).

The 9 items concern **sequence generation** (relevant for selection bias), **allocation concealment** (relevant for selection bias), **confounders** (relevant for selection bias; only for non-randomized studies, i.e. NRCT and NRS), **blinding** (relevant for performance, detection and attrition bias), **incomplete outcome data** (relevant for attrition bias), **selective outcome reporting** (relevant for reporting bias), **other potential threats to validity** (relevant for performance, detection and other sources of bias), **a priori protocol and a priori analysis plan** (relevant for reporting bias).

3.4.3.2 Confounding

An important part of the risk of bias assessment of nonrandomised studies (NRCT and NRS) is how the studies deal with confounding factors (see appendix 7.3). Selection bias is understood as systematic baseline differences between groups and can therefore compromise comparability between groups. Baseline differences can be observable (e.g. age and gender) and unobservable (to the researcher; e.g. motivation and 'ability'). There is no single non-randomised study design that always solves the selection problem. Different designs solve the selection problem under different assumptions and require different types of data. Especially it varies how different designs deal with selection on unobservables. The "right" method depends on the model generating participation, i.e. assumptions about the nature of the process by which participants are selected into a programme.

As there is no universal correct way to construct counterfactuals we will assess the extent to which the identifying assumptions (the assumption that makes it possible to identify the counterfactual) are explained and discussed (preferably the authors should make an effort to justify their choice of method).

For this review, we have identified the following observable confounding factors to be the most relevant: age, gender, education, ethnicity, labour market conditions, censoring and unemployment duration. In each study, we will assess whether these confounding factors have been considered, and in addition we will assess other confounding factors considered in the individual studies. Furthermore, we will assess how each study deals with unobservables.

3.4.3.2.1 Importance of Pre-specified Confounding Factors

The motivation for focusing on age, gender, education and ethnicity is that they are the major determinants of the risk of being unemployed (Layard et al., 2005). Concerning unemployment duration, most studies find that the genuine duration dependence is negative, i.e. the longer the unemployment spell the smaller is the chance of finding a job[7] (see Serneels, 2002, for an overview). If the study does not disentangle the effect of the benefit exhaustion from the negative duration dependence the effect of benefit exhaustion will be biased.

Another potential source of bias is differences in labor market conditions. If the study, for example, explores changes in the duration of benefit entitlement over time or space as the source of variation, it is very important to control for changes in labor market conditions over time (as a consequence of the business cycle, for example) or over space as the exit rate to employment most certainly will depend on this factor.

Censoring may also introduce bias. The effect of effect of benefit exhaustion is often measured with survival data. Participants who do not leave the unemployment system before the end of the study are censored from the outcome data. If not adequately accounted for, such censoring has the potential for introducing bias. Therefore censoring of participants is a potential threat, both in relation to the level of censoring and in relation to whether censoring is taken into account.

3.4.3.3 Assessment

Review authors (at least two, TF & ADK) will independently assess the risk of bias for each included study. Disagreements will be sought by a third reviewer with content and statistical expertise (LP). We will report the risk of bias assessment in risk of bias tables for each included study in the completed review. This assessment will also inform sensitivity analysis.

3.4.4 Measures of Treatment Effect

We expect that the primary treatment effect will be measured either as the impact on the hazard rate or as the impact on the probability of employment at some time interval before the exhaustion of benefits. Our main interest is to include studies in a meta-analysis where hazard ratios and variance are either reported or are calculable from the available data. The hazard rate is defined as the event rate at time *t* conditional on survival until time *t* or later. The incentive effect is the difference in hazard rates *prior* to benefit exhaustion or shortly thereafter between persons who approach exhaustion and persons who do not approach exhaustion (or the hazard rate they *would* have had *if* they were not approaching exhaustion). Depending on the model specification, the incentive effect may be given as a relative change in hazard rates or as an absolute change in hazard rates (van den Berg, 2001; Jenkins,

[7] The reason for this is that unemployment implies a loss of skills or that long periods of unemployment lead to a loss of self-confidence. This “genuine” duration dependence should not be confused with sorting which is another mechanism.

2005; Abbring and van den Berg, 2003). Furthermore, some studies may estimate the incentive effect using indicator variables for the number of months or weeks until exhaustion, whereas other studies may use parametric model specifications. We will allow the outcome measure to be reported in any of these ways.

The effect size will be measured as log hazard ratios (HRs). The log hazard ratio measures the risk of event, in this case the risk of work, in the treatment group in comparison to the control group. Should sufficient effect sizes be available, we will perform a meta-analysis on the individual included studies using the log hazard ratio and variance. We will report the 95% confidence intervals. If log hazard ratios and variances are not reported; and dependent on available data, log hazard ratios and variances will be computed from other information reported in the studies (Parmar, 1998). Log hazard ratios and variances will be computed directly using the observed number of events and logrank expected number of events if available; or indirectly if the p-value for the log-rank, Mantel-Haenszel or chi-squared test if one of these is reported (Sutton et al 2000). Failing this, individual participant data will *not* be requested to calculate log hazard ratios as this may introduce bias due to the time span of studies (the earliest we know of is from 1985 and the latest is from 2009).

The acceptable data sampling frequency is three months or less. Studies using data sampled less frequently than three months will be excluded from the data synthesis. Studies only reporting outcomes measured on time intervals of more than three months will also be excluded from the data synthesis.

The relative risk ratio effect size will be computed for studies only reporting outcomes measured as employment status at some time interval before the exhaustion of benefits or shortly thereafter where calculation of hazard ratios is not possible. Should sufficient data be available for meta-analysis, individual studies will be pooled using the relative risk ratio and variance. We will report the 95% confidence intervals. However, outcomes measured at one time point are inferior to log hazard ratios because the latter takes into account the full observation period, while the former does not (van den Berg, 2001; Jenkins, 2005; Abbring and van den Berg, 2003). In addition, the outcome of point measures might be biased because the researchers might have chosen to report results at time points with significant effects and do not report results at other time points with no difference between groups (selective reporting).

For secondary outcomes, duration of re-employment may be measured as hazard rates in which case the effect size will be measured as log hazard ratios or relative risk ratio. We will report the 95% confidence intervals. Alternatively it may be measured directly as mean duration. Income may be measured as the mean income at different time points or during different time periods. For such continuous outcomes, effect sizes will be calculated if standard deviations are available. Hedges' *g* will be used for estimating standardized mean differences (SMD) where scales

measure the same outcomes in different ways. We will report the 95% confidence intervals.

3.4.5 Unit of Analysis Issues

We will take into account the unit of analysis of the studies to determine whether individuals were randomised in groups (i.e. cluster randomised trials), whether individuals may have undergone multiple interventions, whether there were multiple treatment groups and whether several studies are based on the same data source.

3.4.5.1 Multiple Intervention Groups

Multiple intervention groups (with different individuals) within a study with one control group will not be pooled, nor will multiple controls groups be pooled. Data will be rigorously checked to avoid overlapping samples in the meta-analysis.

3.4.5.2 Multiple Interventions per Individual

Only one intervention group will be coded and compared to the control group to avoid overlapping samples. An example may be individuals who exit to other types of time limited unemployment benefits when unemployment insurance benefits exhaust. The intervention included will be the first exhaustion experienced by the individual. Data from studies comparing different types of interventions/comparisons will be coded and analysed separately.

3.4.5.3 Multiple Studies using the Same Sample of Data

In some cases, several studies may have used the same sample of data (e.g. studies using the same register data). We will review all such studies, but in the meta-analysis we will only include one estimate of the benefit exhaustion effect from each sample of data. This will be done to avoid dependencies between the "observations" (i.e. the estimates of the benefit exhaustion effect) in the meta-analysis. The choice of which estimate to include will be based on our quality assessment of the studies. We will choose the estimate from the study that we judge to have the least risk of bias and the judgement will pay particular attention to selection bias.

3.4.5.4 Multiple Time Points

When the results are measured at multiple time points, each outcome at each time point will be analyzed in a separate meta-analysis with other comparable studies taking measures at a similar time point. As a guideline these will be grouped together as follows: short-term (-1-1 months before/after benefit exhaustion), medium term (1-2 months before benefit exhaustion) long term (at least 2 months before benefit exhaustion). However, should the studies provide viable reasons for adjusted choice of relevant and meaningful duration intervals for the analysis of outcomes, we will adjust the grouping.

3.4.5.5 Cluster Randomisation

In cluster randomisation statistical analysis errors can occur when the unit of allocation is different from the unit of analysis. In cluster-randomised trials, the elements are groups of individuals (e.g. region, PES office), rather than individuals themselves. In such studies, care should be taken to avoid biased standard errors if the unit-of-analysis is the individual. When suitable cluster analysis is used, effect estimates and their standard errors will be meta-analysed (Higgins & Green, 2008). In cases where study investors have not applied appropriate analysis methods controlling for clustering, we will try to measure the intra-cluster correlation to correct standard errors (see Donner, 2001).

3.4.6 Dealing with Missing Data and Incomplete Data

Missing data, attrition rates and censoring will be assessed in the included studies. Notably due to the nature of the field studies of benefit exhaustion typically estimate the effect on data collected from administrative registers or by questionnaires. Especially studies using data collected from questionnaires are subject to missing data. For studies using questionnaire data, a sensitivity analysis will be used to assess potential bias, and the extent to which the results might be biased by missing data will be discussed. The review authors will record attrition rates and (if possible) reasons for attrition from included studies.For studies in which the censoring level is high (more than 25%) or the level is not reported, a sensitivity analysis will be used to assess potential bias in the analysis and the extent to which the results might be biased by a high censoring level will be discussed.

Information on intention to treat analysis (ITT) will also be recorded for RCTs and QRCTs. As we expect the majority of studies to be non-randomised (NRCTs and NRSs) and treatment on the treated (TOT) analysis used sensitivity analysis will be run to examine influences on the effects of controlled trials where adequate ITT analysis are used.

3.4.7 Assessment of Heterogeneity

Statistically significant heterogeneity among primary outcome studies will be assessed with Chi-squared (Q) test and I-squared (Higgins et al., 2003). A significant Q ($p<.05$) and I-squared of at least 50% will be considered as statistical heterogeneity.

3.4.8 Assessment of Reporting Bias

Reporting bias refers to both publication bias and selective reporting of outcome data and results. Here we state how we will assess publication bias. We will use funnel plots for information about possible publication bias if we find sufficient studies (Higgins & Green, 2008). However asymmetric funnel plots are not necessarily caused by publication bias (and publication bias does not necessarily cause asymmetry in a funnel plot). If asymmetry is present, we will consider possible reasons for this.

3.5 DATA SYNTHESIS

Studies that have been coded with a very high risk of bias (5 on the risk of bias scale) will not be included in the data synthesis. All follow-up durations reported in the primary studies will be recorded and we will do separate analyses for short-term, medium-term and long-term outcomes. Otherwise included studies will be pooled where appropriate dependent on heterogeneity in relation to covariates such as participant characteristics (age, gender, education and ethnicity), labour market conditions, type of unemployment benefit, duration of entitlement and if compulsory activation is part of the system.

As we expect the studies to deal with diverse populations of participants and labour market conditions, the choice of a random effects model to represent the overall effect seems to be the most adequate option. For subsequent analyses of moderator variables in search for systematic variations we will switch to the mixed model (if a predictor that explains some between-studies variation is available but we also want to account for the remaining uncertainty, the mixed-effects regression model is appropriate (Hedges and Pigott, 2004; Konstantopoulos, 2006)

3.5.1 Moderator Analysis and Investigation of Heterogeneity

If heterogeneity is judged to be large we will investigate the following factors (if they are available) with the aim of explaining observed heterogeneity: study-level summaries of participant characteristics (e.g. studies considering a specific age group, gender or educational level or studies where separate effects for men/women, young/old or low/high educational level are available), labour market conditions (good/bad), type of unemployment benefit (UI or SA/UA), duration of entitlement (less than one year , between one and two years, more than two years), whether alternative benefits are available and if compulsory activation is part of the system.

If the number of included studies is sufficient and given there is variation in the covariates, we will perform moderator analyses (meta-regression using the mixed model) to explore how observed variables are related to heterogeneity. We will estimate the (new) residual variance component to be used in a weighted least squares analysis conditional on this variance component estimate. The residual variance component will be estimated using the method-of-moments estimator (Hartung et al., 2008 and Konstantopoulos, 2006). We will report the 95% confidence intervals for regression parameters. Conclusions from meta-regression analysis will be cautiously drawn and will not be based on significance tests.

Otherwise single factor subgroup analysis will be performed. The assessment of any difference between subgroups will be based on 95% confidence intervals. No conclusions from subgroup analyses will be drawn and interpretation of relationships will be cautious, as they are based on subdivision of studies and indirect comparisons.

3.5.2 Sensitivity Analysis

Sensitivity analysis will be used to evaluate whether the pooled effect sizes are robust across study design and components of methodological quality. For methodological quality, we will consider sensitivity analysis for each major component of the risk of bias checklists. Sensitivity analysis will further be used to examine the robustness of conclusions in relation to the quality of data (outcome measures based on different time intervals and whether data is based on questionnaires or administrative registers).

3.6 TIMEFRAME

Review authors intend to complete this work within a year from protocol approval and to update within three years from review approval.

4 Sources of support

4.1 INTERNAL SOURCES

SFI Campbell, Copenhagen, Denmark.
Review team at SFI Campbell consists of research assistants Anne-Sofie Due Knudsen, Simon Helth Filges and Stine Lian Olsen.

4.2 EXTERNAL SOURCES

5 References

Abbring, J.H., & van den Berg, G.J. (2003). The nonparametric identification of treatment effects in duration models. *Econometrica, 71*(5) (September), 1491-1517.

Acemoglu, D., & Shimer, R. (1999). Efficient unemployment insurance. *Journal of Political Economy, 107,* 893-928.

Boone, J., & J.C. van Ours. (2009). Why is there a spike in the job finding rate at benefit exhaustion? IZA DP No. 4523.

Card, D., Chetty, R., & Weber, A. (2007). The spike at benefit exhaustion: Leaving the unemployment system or starting a new job. IZA DP No. 2590.

Caliendo, M., Tatsiramos, K., & Uhlendoff, A. (2009). Benefit duration, unemployment duration and job match quality: A regression-discontinuity approach. IZA DP No. 4670.

Donner, A., Piaggio, G., & Villar, J. (2001). Statistical methods for the meta-analysis of cluster randomized trials. *Statistical Methods in Medical Research, 10*, 325-338.

Feldstein, M. (2005). Rethinking social insurance. *The American Economic Review, 95* (1), 1-24.

Fredriksson, P., & Holmlund, B. (2006). Improving incentives in unemployment insurance: A review of recent research. *Journal of Economic Surveys, 20(3*).

Geerdsen, L.P. (2002). The identification of incentive effects of benefit exhaustion in unemployment insurance systems. (Working Paper 25:2002). The National Institute of Social Research, SFI.

Geerdsen, L. P., & Holm, A. (2007). Duration of UI periods and the perceived threat effect from labour market programmes. Labour Economics, 14,639-652.

Geerdsen, L.P., Bjørn, N.H., Filges, T., & Jensen, P. (2011). The Threat of Active Labour Market Programmes for Unemployed. A systematic review submitted to the Campbell Collaboration.

Gustafson, C., & Levine, P. (1997). Less-skilled workers, welfare reform and the unemployment insurance. Working Paper. University of California at Berkeley, Dept. of Economic.

Hedges, L. V. (2007). Meta-analysis. In C.R. Rao (Ed.), The handbook of statistics (pp 919-53). Amsterdam: Elsevier.

Hartung, J., Knapp, G., & Sinha, B. K. (2008). Statistical meta-analysis with applications. New York: Wiley.

Hedges, L.W., & Pigott, T.D. (2004). The power of statistical tests for moderators in meta-analysis. Psychological Methods, 9(4), 426–445.

Higgins, J.P., Thompson, S.G., Deeks, J.J., & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. British Medical Journal, 327(7414), 557-560.

Higgins, J.P.T., & Green, S. T. (Eds.). (2008). Cochrane handbook for systematic reviews of interventions. Chichester, UK:Wiley-Blackwell.

Jackman R., & Layard, R. (1991) Does long-term unemployment reduce a person's chance of a job? A time series test. Economica, 58, 93–106.

Jenkins, S. P. (2005), Survival Analysis. Lecture Notes, 18 July. Retreived from http://www.iser.essex.ac.uk/files/teaching/stephenj/ec968/pdfs/ec968lnotesv6.pdf

Kalwij, A. (2010). Unemployment durations and the pattern of duration dependence over the business cycle of British males. Empirical Economics, 38, 429–456.

Katz, L. F., & Meyer, B.D. (1990). The impact of the potential duration of unemployment benefits on the duration of unemployment. Journal of Public Economics, 41,45-71. Amsterdam: North-Holland.

Kluve, J., Lehmann, H., & Schmidt, C.M. (1999). Active labor market policies in Poland: Human capital enhancement, stigmatization or benefit churning. Journal of Comparative Economics, 27, 61-68.

Kluve, J., Card, D., Fertig, M., Góra, M., Jacobi, L., Jensen, P., . . . Weber, A. (2007). Active labor market policies in Europe: Performance and perspectives. Berlin: Springer Verlag.

Konstantopoulos, S. (2006). Fixed and mixed effects models in meta-analysis. IZA DP no. 2198.

Lalive, R., van Ours, J., & Zweimüller, J. (2006). How changes in financial incentives affect the duration of unemployment. Review of Economic Studies, 73(4), 1009-1038. Oxford, UK: Oxford University Press.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. Econometrica, 47(4), 939–956.

Layard, R., Nickell, S., & Jackman, R. (2005). Unemployment. Oxford, UK: Oxford University Press.

Parmar, M.K.B., Torri, V., & Stewart L. (1998). Extracting summary statistics to perform meta analyses of the published literature for survival endpoints. Statistics in Medicine, 17, 2815-2834.

Marimon, R., & Zilibotti, F. (1999). Unemployment vs. mismatch of talents: Reconsidering unemployment benefits. Economic Journal, 109,266–291.

Meyer, B. D. (1990). Unemployment insurance and unemployment spells. Econometrica, 58(4), 757-782.

Mortensen, D. (1977). Unemployment insurance and job search decisions. Industrial and Labor Relations Review, 30, 505–517.

Mortensen, D. T. (1987). Job search and labor market analysis. In O. Ashenfelter & R. Layard (Eds.), Handbook of labor economics (Vol. 2). Amsterdam: North-Holland.

Narendranathan, W., & Stewart, M.B. (1993) Modelling the probability of leaving unemployment: competing risks models with flexible baseline hazards. Journal of the Royal Statistical Society. Series C (Applied Statistics), 42(1),63-83.

Nickell, S., Nunziata, L, & Ochel, W. (2005). Unemployment in the OECD since the 1960s. What do we know? The Economic Journal, 115, 1-27.

Salant, S.W. (1977): Search theory and duration data: a theory of sorts. Quarterly Journal of Economics, 91(1),39–57.

Serneels, P. (2002). Explaining Non-Negative Duration Dependence Among the Unemployed. CSAE Working Paper Series 2002-13, Oxford, UK: Centre for the Study of African Economies, University of Oxford.

Simmonds, M.C., & Higgins, J.P.T. (2007). Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. Statistics in Medicine, 26, 2982-2999.

Solow, R. (2000). Unemployment in the United States and in Europe – A contrast and the reasons. CESifo Working Paper Series 231. Munich: CESifo Group Munich

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). Methods of meta-analysis in medical research. New York: Wiley.

Van den Berg, G.J. (1990). Nonstationarity in job search rheory. Review of Economic Studies, 57,255–277.

Van den Berg, G.J. (2001). Duration Models: Specification, Identification, and Multiple Durations. In J.J. Heckman & E. Leamer (Eds.), Handbook of econometrics, Volume V. Amsterdam: North-Holland.

6 Appendices

6.1 FIRST AND SECOND LEVEL SCREENING

First level screening is on the basis of titles and abstracts. Second level is on the basis of full text

Reference id. No. :
Study id. No.:
Reviewers initials:
Source:
Year of publication:
Duration of study:
Country/countries of origin
Author

The study is excluded if one or more of the answers to question 1-3 are No. If the answers to question 1 to 3 are yes or uncertain then the full study is retrieved for second level eligibility. All uncertain questions need to be posed again on the basis of fill text. If not enough information is available or if the study is unclear the author of the study will be contacted if possible.

First level screening questions are based on titles and abstracts

1. Are the participants' unemployed individuals receiving some kind of benefit during their unemployment?
   Yes - include
   No – if no then stop here and exclude
   Uncertain - include

Question 1 guidance:
This includes all types of unemployment benefits both unemployment insurance benefits, unemployment assistance benefits and social assistance benefits.

2. Does the study focus on time limits in the unemployment benefit eligibility period or exhaustion of unemployment benefits or entitlement to

unemployment benefits or maximum duration of unemployment benefits etc.?
Yes - include
No – if no then stop here and exclude
Uncertain - include

Question 2 guidance:
The intervention is the exhaustion of any kind of unemployment benefit with a known expiration date. This intervention can be referred to in different ways.

3. Is this study a primary quantitative study?
   Yes - include
   No – if no then stop here and exclude
   Uncertain - include

Question 3 guidance:
We are only interested in primary quantitative studies, where the authors have analyzed the data. We are not interested in theoretical papers on the topic or surveys/reviews of studies of the topic. (This question may be difficult to answer on the base of titles and abstracts alone.)

Second level screening questions based on full text

4. Does the study estimate an effect, using a control group or using an estimated counterfactual?
   Yes - include
   No – if no then stop here and exclude
   Uncertain - include

Question 4 guidance
E.g. 1) Randomised controlled trials including cluster randomisation and quasi randomised controlled study designs (i.e. participants are allocated by means such as alternate allocation, person's birth date, the date of the week or month, case number or alphabetical order), 2) non randomised controlled study designs (i.e. quasi-experimental designs) such as controlled two group study designs or 3) study designs based on observational data, where the effect is estimated by statistical methods.

5. Does the study examine exits to employment?
   Yes – include
   No – if no then stop here and exclude
   Uncertain – include

Question 3 guidance:
The primary outcome is exits to employment. Studies only looking at exits to other destinations (such as other kinds of benefits or out of the labor force) or studies who do not distinguish between destinations will not be included.

6.2 DATA EXTRACTION

Language
Journal
Year
Country
Time period covered by data
Type of unemployment scheme (UI, social benefit other (specify))
Target group (age, gender, education, eligibility requirements for benefits)
Benefit level/replacement rate
Labour market conditions (unemployment rate)
Benefit level/replacement rate available after exhaustion if any
Is compulsory activation part of the system? If yes, describe the elements of the programme (education, work, training, self-employment, job search assistance)
Maximum duration of unemployment benefits
Type of data used (register, questionnaire, other (specify))
Sampling frequency
Time interval the outcome measure is based on (if different from sampling frequency)
Is there correction for unobserved heterogeneity? If yes, how?
Sample size (Treatment/control)
Censoring level (percent, separate for intervention/control if possible)
Is there correction for censoring (yes/no)

Outcome measures

Instructions: Please enter outcome measures in the order in which they are described in the report. Note that a single outcome measure can be completed by multiple sources and at multiple points in time (data from specific sources and time-points will be entered later).

#	Outcome & measure	Reliability & Validity	Format	Direction	Source	Blind (outcome assessors) ?	Pg# & notes
1		Info from: Other samples This sample Unclear Info provided:	Dichotomy Continuous	High score or event is Positive Negative Can't tell	Questionnaire Admin data Other (specify) Unclear	Yes No Can't tell	

* Repeat as needed

OUT COME DATA

DICHOTOMOUS OUTCOME DATA

OUTCOME	TIME POINT (s) (record exact time from exhaustion, there may be more than one, record them all)	SOURCE	VALID Ns	CASES	NON-CASES	STATISTICS	Pg. # & NOTES
		Questionnaire Admin data Other (specify) Unclear	Exhaustion Comparison	Exhaustion Comparison	Exhaustion Comparison	RR (risk ratio) OR (odds ratio) SE (standard error) 95% CI DF P- value (enter exact p value if available) Chi2 Other Covariates (control variables, age, gender, education, ethnicity, duration dependence, labor market conditions, censoring, other)	

Repeat as needed

OUT COME DATA

TIME-TO-EVENT OUTCOME DATA

OUTCOME	TIME POINT (s) (record exact time from exhaustion, there may be more than one, record them all)	SOURCE	Method of estimation			STATISTICS	Pg. # & NOTES
		Questionnaire Admin data Other (specify) Unclear	Non-parametric Semi-parametric Parametric			HR (hazard ratio) SE (standard error) 95% CI DF P- value (enter exact p value if available) Chi2 Other Covariates (control variables, age, gender, education, ethnicity, duration dependence, labor market conditions, censoring, other)	

Repeat as needed

CONTINUOUS OUTCOME DATA

OUTCOME	TIME POINT (s) (record exact time from exhaustion, there may be more than one, record them all)	SOURCE (specify)	VALID Ns	Means	SDs	STATISTICS	Pg. # & NOTES
		Questionnaire Admin data Other (specify) Unclear	Exhaustion	Exhaustion	Exhaustion	P t F Df ES Covariates Other	
			Comparison	Comparison	Comparison		

*Repeat as need

6.3 ASSESSMENT OF RISK OF BIAS IN INCLUDED STUDIES

Risk of bias table

Item	**Judgement**[a]	**Description** (quote from paper, or describe key information)
1. Sequence generation		
2. Allocation concealment		
3. Confounding[b,c]		
4. Blinding?[b]		
5. Incomplete outcome data addressed?[b]		
6. Free of selective reporting?[b]		
7. Free of other bias?		
8. A priori protocol?[d]		
9. A priori analysis plan?[e]		

[a] Some items on low/high risk/unclear scale (double-line border), some on 5 point scale/unclear (single line border), some on yes/no/unclear scale (dashed border). For all items, record "unclear" if inadequate reporting prevents a judgement being made.

[b] For each outcome in the study.

[c] This item is only used for NRCTs and NRSs. It is based on list of confounders considered important at the outset and defined in the protocol for the review (*assessment against worksheet*).

[d] Did the researchers write a protocol defining the study population, intervention and comparator, primary and other outcomes, data collection methods, etc. in advance of starting the study?

[e] Did the researchers have an analysis plan defining the primary and other outcomes, statistical methods, subgroup analyses, etc. in advance of starting the study?

Risk of bias tool

Studies for which RoB tool is intended

The risk of bias model is developed by Prof. Barnaby Reeves in association with the Cochrane Non-Randomised Studies Methods Group.[8] This model, an extension of the Cochrane Collaboration's risk of bias tool, covers both risk of bias in randomised controlled trials (RCTs and QRCTs), but also risk of bias in non-randomised studies (NRCTs and NRSs).

The point of departure for the risk of bias model is the Cochrane Handbook for Systematic Reviews of interventions (Higgins & Green, 2008). The existing Cochrane risk of bias tool needs elaboration when assessing non-randomised studies because, for non-randomised studies, particular attention should be paid to selection bias / risk of confounding. Additional item on confounding is used only for non-randomised studies (NRCTs and NRSs) and is not used for randomised controlled trials (RCTs and QRCTs).

Assessment of risk of bias

Issues when using modified RoB tool to assess included non-randomised studies:

- Use existing principle: score judgment and provide information (preferably direct quote) to support judgment
- Additional item on confounding used only for non-randomised studies (NRCTs and NRSs).
- 5-point scale for some items (distinguish "unclear" from intermediate risk of bias).
- Keep in mind the general philosophy – assessment is not about whether researchers could have done better but about risk of bias; the assessment tool must be used in a standard way whatever the difficulty / circumstances of investigating the research question of interest and whatever the study design used.
- Anchors: "1/No/low risk" of bias should correspond to a high quality RCT. "5/high risk" of bias should correspond to a risk of bias that means the findings should not be considered (too risky, too much bias, more likely to mislead than inform)

1. Sequence generation
   - Low/high/unclear RoB item
   - Always high RoB (not random) for a non-randomised study
   - Might argue that this item redundant for NRS since always high – but important to include in RoB table ('level playing field' argument)

2. Allocation concealment
   - Low/high/unclear RoB item
   - Potentially low RoB for a non-randomised study, e.g. quasi-randomised (so high RoB to sequence generation) but concealed (reviewer judges that the people making decisions about including participants didn't know how allocation was being done, e.g. odd/even date of birth/hospital number)

3. RoB from confounding (additional item for NRCT and NRS; assess for each outcome)
   - Assumes a pre-specified list of potential confounders defined in the protocol
   - Low(1) / 2 / 3 / 4 / high(5) / unclear RoB item

[8] This risk of bias model was introduced by Prof. Reeves at a workshop on risk of bias in non-randomised studies at SFI Campbell, February 2011. The model is a further development of work carried out in the Cochrane Non-Randomised Studies Method Group (NRSMG).

- Judgment needs to factor in:
    - proportion of confounders (from pre-specified list) that were considered
    - whether most important confounders (from pre-specified list) were considered
    - resolution/precision with which confounders were measured
    - extent of imbalance between groups at baseline
    - care with which adjustment was done (typically a judgment about the statistical modeling carried out by authors)
- Low RoB requires that all important confounders are balanced at baseline (not primarily/not only a statistical judgment OR measured 'well' and 'carefully' controlled for in the analysis.

Assess against pre-specified worksheet. Reviewers will make a RoB judgment about each factor first and then 'eyeball' these for the judgment RoB table.

4. RoB from lack of blinding (assess for each outcome, as per existing RoB tool)
- Low(1) / 2 / 3 / 4 / high(5) / unclear RoB item
- Judgment needs to factor in:
    - nature of outcome (subjective / objective; source of information)
    - who was / was not blinded and the risk that those who were not blinded could introduce performance or detection bias
    - see Ch.8

5. RoB from incomplete outcome data (assess for each outcome, as per existing RoB tool)
- Low(1) / 2 / 3 / 4 / high(5) / unclear RoB item
- Judgment needs to factor in:
    - reasons for missing data
    - whether amount of missing data balanced across groups, with similar reasons
    - whether censoring is less than or equal to 25% and taken into account
    - see Ch.8

6. RoB from selective reporting (assess for each outcome, NB different to existing Ch.8 recommendation)
- Low(1) / 2 / 3 / 4 / high(5) /unclear RoB item
- Judgment needs to factor in:
    - existing RoB guidance on selective outcome reporting
    - see Ch.8
    - also, extent to which analyses (and potentially other choices) could have been manipulated to bias the findings reported, e.g. choice of method of model fitting, potential confounders considered / included
    - look for evidence that there was a protocol in advance of doing any analysis / obtaining the data (difficult unless explicitly reported); NRS very different from RCTs. RCTs must have a protocol in advance of starting to recruit (for REC/IRB/other regulatory approval); NRS need not (especially older studies)
    - Hence, separate yes/no items asking reviewers whether they think the researchers had a pre-specified protocol and analysis plan.

Confounding Worksheet

Assessment of how researchers dealt with confounding	
Method for *identifying* relevant confounders described by researchers: yes no If yes, describe the method used:	☐ ☐
Relevant confounders described: yes no List confounders described on next page	☐ ☐
Method used for controlling for confounding At design stage (e.g. matching, regression discontinuity, instrument variable): .... At analysis stage (e.g. stratification, multivariate regression, difference-indifference): .... Describe confounders controlled for below	

Confounders described by researchers

Tick (yes[0]/no[1] judgment) if confounder considered by the researchers [Cons'd?]

Score (1[good precision] to 5[poor precision]) precision with which confounder measured

Score (1[balanced] to 5[major imbalance]) imbalance between groups

Score (1[very careful] to 5[not at all careful]) care with which adjustment for confounder was carried out

Confounder	Considered	Precision	Imbalance	Adjustment
Gender	☐	☐	☐	☐
Age	☐	☐	☐	☐
Ethnicity	☐	☐	☐	☐
Education	☐	☐	☐	☐
Labour market condition	☐	☐	☐	☐
Unemployment duration	☐	☐	☐	☐
Unobservables[9]	☐	Irrelevant	☐	☐
Censoring	☐	☐	☐	☐
Other:	☐	☐	☐	☐

[9] See user guide for unobservables

User guide for unobservables

Selection bias is understood as systematic baseline differences between groups and can therefore compromise comparability between groups. Baseline differences can be observable (e.g. age and gender) and unobservable (to the researcher; e.g. motivation and 'ability'). There is no single non-randomised study design that always solves the selection problem. Different designs solve the selection problem under different assumptions and require different types of data. Especially how different designs deal with selection on unobservables varies. The "right" method depends on the model generating participation, i.e. assumptions about the nature of the process by which participants are selected into a programme.

As there is no universal correct way to construct counterfactuals we will assess the extent to which the identifying assumptions (the assumption that makes it possible to identify the counterfactual) are explained and discussed (preferably the authors should make an effort to justify their choice of method). We will look for evidence that authors using e.g. (this is NOT an exhaustable list):

Natural experiments:
Discuss whether they face a truly random allocation of participants and that there is no change of behavior in anticipation of e.g. policy rules.

Instrument variable (IV):
Explain and discuss the assumption that the instrument variable does not affect outcomes other than through their effect on participation.

Matching (including propensity scores):
Explain and discuss the assumption that there is no selection on unobservables, only selection on observables.

(Multivariate) Regression:
Explain and discuss the assumption that there is no selection on unobservables, only selection on observables. Further discuss the extent to which they compare comparable people.

Regression Discontinuity (RD):
Explain and discuss the assumption that there is a (strict!) RD treatment rule. It must not be changeable by the agent in an effort to obtain or avoid treatment. Continuity in the expected impact at the discontinuity is required.

Difference-in-difference (Treatment-control-before-after):
Explain and discuss the assumption that outcomes of participants and nonparticipants evolve over time in the same way.