Online appendix C: Risk of bias assessment tool

The following table provides a provisional tool to guide the risk of bias assessment for quantitative impact evaluations. If necessary, we could amend the tool to better inform the appraisal of primary studies.

Provisional risk of bias assessment tool (RCT)

General	ID	EPPI ID		
General	Study first author	Open answer		
General	Time taken to complete assessment	Minutes		
General	Design type: What type of study design is used?	1= Randomised controlled trial (RCT) (random assignment to households/individuals) or quasi-RCT 2= Cluster-RCT (quasi-RCT)	-	
General	Methods used for analysis: Which methods are used to control for selection bias and confounding?	1 = Statistical matching (PSM, CEM, covariate matching) 2 = Difference in differences (DID) estimation methods 3 = IV-regression (2-stage least squares or bivariate probit) 4 = Heckman selection model 5 = Fixed effects regression 6 = Covariate adjusted estimation 7 = Propensity weighted regression 8 = Comparison of means 9 = Other (please state)	-	
General	Design and analysis method description	Open answer	Briefly describe the study design and analysis method undertaken by the authors.	

General	Study population	Open answer	Provide any details in the paper that describe how the study population was selected, covering: a) How is the population selected? what is the sampling strategy to recruit participants from that population into the study? b) What are the characteristics of that study participants? c) Was this a pilot program aimed at being scaled up? d) Were there specific factors of success or failure in the implementation?	
General	Type of comparison group	1=No intervention (service delivery as usual) 2=Other intervention 3=Pipeline (wait-list) control (still service delivery as usual)	Indicate type of comparison group	
General	Type of comparison group (if other)	Open answer		
General	Ethical clearance	Open answer	Provide any details of ethical research clearances granted. Report unclear if this information is not available.	
General	Study registration	Open answer	Provide any detials of study registration, inlcuding registry IDs, etc.	
1: Assignment mechanism - Assessment	Assignment mechanism: Was the allocation or identification mechanism random or as good as random?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) The authors describe a random component in sequence generation/ randomization method (e.g. lottery, coin toss, random number generator) and assignment is performed for all units at the start of the study centrally or using a method concealed from participants and intervention delivery b) If public lottery is used for the sequence	Score “Yes” if all criterion a), b), c) and d) are satisfied. Score "Probably Yes" if only criterion a) and b) are not satisfied OR if only criteria c) is not satisfied. Score “Unclear” if d) is not satisfied because no balance table is

			generation, authors provide detail on the exact settings and participants attending the lottery. c) If a special randomization procedure is used to ensure balance, it is well described and justified given the study setting (stratification, pairwise matching, unique random draw, multiple random draws etc). d) A balance table is reported suggesting that allocation was random between all groups including subgroup receiving different treatment within control or treatment groups (if the comparison is relevant for this assessment).	reported. Score "Probably No" if d) is not satisfied because there is no balance table reported and there is evidence suggesting a problem in the randomization, such as baseline coefficients in a diff-in-diff regression table are very different or sample size is too small for the procedure used (using stratification when there are less than two units for each intervention and control group in each strata can lead to imbalance). Score “No” if d) is not satisfied because there are large imbalances concerning a large number of variables, providing evidence that the assignment was not random. If this is scored as no, use the NRS tool.
1: Assignment mechanism - Justification	Assignment justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
2: Unit of analysis - Assessment	Unit of analysis: Is unit of analysis in cluster allocation addressed in standard error calculation ?	1=Yes 2=No 3=Not reported/unclear 4=Not applicable	Score "Yes" if UoA = UoR OR if UoA != UoR and standard errors are clustered at the UoR level OR data is collapsed to the UoR level Score "Not reported/unclear" if not enough information is provided on the way the standard errors were calculated or what the	

			unit of analysis is. Score "Not applicable" if it is not a cluster RCT. Score "No" otherwise.	
2: Unit of analysis - Justification	Method used to address differences between UoA and unit of data collection	Open answer		
3: Selection bias - Assessment	Selection bias Was any differential selection into or out of the study (attrition bias) adequately resolved?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	Score "Yes" if there is no attrition or attrition falls into the green zone and the study establishes that attrition is randomly distributed (e.g. by presenting balance by key characteristics across groups) AND if survey respondents were randomly sampled. Score "Probably yes" if attrition falls into the green zone AND if survey respondents were randomly sampled. Score "Unclear" if there is an attrition problem but no information provided on the relationship between attrition and treatment status, OR if there is not enough information on how the population surveyed was sampled. Score "Probably no" if there is attrition which is likely to be related to the intervention OR there is some indication that the survey respondents were purposely	

			sampled in a way that might have led the sampling to be different between treatment and control groups, or attrition falls into the yellow zone. Score "No" if attrition falls into the red zone.	
3: Selection bias - Justification	Selection bias justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
3: Confounding - Assessment	Confounding and group equivalence: Was the method of analysis executed adequately to ensure comparability of groups throughout the study and prevent confounding	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) Baseline characteristics are similar in magnitude; b) Unbalanced covariates at the individual and cluster level are controlled in adjusted analysis; c) Adjustments to the randomization were taken into account in the analysis (stratum fixed effects, pairwise matching variables)? (Bruhn and McKenzie 2009)	Score “Yes” if criterion a) and b) are satisfied; Score "Probably yes" if a) is not satisfied but b) is satisfied and imbalances are small in magnitude OR if only a) is satisfied. Score “Unclear” if no balance table is provided or if imbalances are controlled for but they are very large in magnitude and assignment mechanism is not coded as "Yes" or "Probably yes" Score "Probably no" if a) and b) are not satisfied and the magnitude of imbalances are small Score “No” if a) and b) are not satisfie and the magnitude of imbalances are large and covariates are clear determinant of the outcomes.

3: Confounding - Justification	Confounding justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	

4: Deviations from intended interventions - Assessment	Deviations from intended interventions: Spill-overs, cross-overs and contamination: was the study adequately protected against spill-overs, cross-overs and contamination?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) There was no implementation issues that might have led the control participants to receive the treatment (implementer's mistake). b) The intervention is unlikely to spill-over to comparisons (e.g. participants and non-participants are geographically and/or socially separated from one another and general equilibrium effects are not likely) or the potential effects of spill overs were measured (e.g. variation in the % of unit within a cluster receiving the treatment). c) There is no risk of contamination by external programs: the treatment and comparisons are isolated from other interventions which might explain changes in outcomes. d) There is nothing in the surveys that might have given the control participants an idea of what the other group might receive OR they did but there is no risk that this has changed their behaviours; AND the survey process did not reveal information to the control group that they did not have before (e.g. the study aims to measure increase in take up of a service or product that participants might not know about) Authors might put something in place in the design of the study that allows to control for that survey effect (e.g. a pure control with no monitoring except baseline end line)	Score “Yes” if criterion a), b), c) and d) are satisfied; Score "Probably yes" if there is no obvious problem but there is no information reported on potential risks related to spill overs, contamination, or survey effects in the control group OR if there were issues with spill-overs but they were controlled for or measured. Score “Unclear” if spill-overs, cross-overs, survey effects and/or contamination are not addressed clearly. Score "Probably no" if any of the criterion a), b), c) or d) are not satisfied but the scale of the issue is not clear. Score “No” if any of the criterion a), b), c) or d) are not satisfied and happened at a large scale in the study.

4: Deviations from intended interventions - Justification	Deviations justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages). For example, intervention groups are geographically separated, authors use intention to treat estimation or instrumental variables to account for non-adherence, and survey questions are not likely to expose individuals in the control group to information about desirable behaviours ('survey effects').	
5. Performance bias - Assessment	Performance bias: Was the process of monitoring individuals unlikely to introduce motivation bias among participants?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) The authors state explicitly that the process of monitoring the intervention and outcome measurement is blinded and conducted in the same frequency for treatment and control groups, or argue convincingly why it is not likely that being monitored could affect the performance of participants in treatment and comparison groups in different ways (such as resulting in Hawthorne or John Henry effects). b) The outcome is based on data collected in the context of a survey, and not associated with a particular intervention trial, or data are collected from administrative records or in the context of a retrospective (ex post) evaluation.	Score "Yes" if either criterion a) or b) are satisfied; Score "Probably yes" if the study is based on data collected during a trial and there is no obvious issue with the monitoring processes but authors do not mention potential risks. Score "Unclear" if it is not clear whether the authors use an appropriate method to prevent Hawthorne and John Henry Effects (e.g. blinding of outcomes and, or enumerators, other methods to ensure consistent monitoring across groups). Hawthorne effects may result where participants know that they are being observed and John Henry Effects may result from participant knowledge of being

				compared. Score "Probably no" if there was imbalance in the frequency of monitoring in intervention groups, which might have influenced participants' behaviours. Score "No" if neither criterion a) or b) are satisfied.
5. Performance bias - Justification	Performance bias justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	

6. Outcome measurement bias - Assessment	Outcome measurement bias: Was the study free from biases in outcome measurement?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) Outcome assessors are blinded or the outcome measures are not likely to be biased by their judgement. b) For self-reported outcomes: respondents in the intervention group are not more likely to have accurate answers due to recall bias; c) For self-reported outcomes: respondents do not have incentives to over/under report something related to their performance or actions, OR researchers put in place mechanisms to reduce the risk of reporting bias (researchers not strongly involved in the implementation of the program and it is clear that their answers to the survey will not affect what they receive in the future) OR authors have measured the risks of bias through falsification tests or measuring the effect on placebo outcomes in cases where there was a risk of reporting bias. d) Timing issue: the data collection period did not differ between intervention and comparison group, the baseline data is not likely to be affected by the beginning of the intervention or affects a small percentage of the study participants.	Score “Yes” if criterion a), b), c) and d) are satisfied: Score "Probably yes" if there is a small risk related to any of a), b), c) or d) and there is no more information provided to to justify the absence of bias OR if there was a high risk of bias but authors have either controlled it in their design or measured it with a placebo outcomes. Score “Unclear” if it there is a high risk related to any of a), b), c) or d) and there is no more information provided to to justify the absence of bias. Score "Probably no" if there are high risk related to a), b), c) or d) and it is clear that authors were not able to control for this bias. Score “No” if there is evidence of bias.
6. Outcome measurement bias - Justification	Outcome measurement justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	

7. Reporting bias - Assessment	Analysis reporting: Was the study free from selective analysis reporting?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) A pre-analysis plan or trial protocol is published and referred to or the trial was pre-registered or the outcomes were pre-registered; b) Authors report results corresponding to the outcomes announced in the method section (there is no outcome reporting bias); c) Authors report results of unadjusted analysis and intention to treat (ITT) estimation, alongside any adjusted and treatment-on-the-treated/complier-average-causal-effects analysis.) d) Authors use the appropriate analysis method (use baseline data when available) and different treatment arms are differentiated in the analysis e) Authors have reported all the analysis which could help understand the results and no other bias is assessed as unclear due to the lack of an important analysis (e.g. a balance table or a subgroup analysis)	Score "Yes" if all the criterion a), b), c), d), and e) are satisfied; Score "Probably yes" if all the conditions are met except a), or if all the conditions are met but there is some element missing that could have helped understand the results better (e); Score "Unclear" if there is not enough information to determine that there is an analysis missing; Score "Probably no" if any of the criterion b), c) or d) are not satisfied; Score "No" if any of the criterion b), c) or d) are not satisfied and there is evidence that the analysis results would be different because large imbalances were not controlled for, compliance was very low and ITT estimation was not reported or different treatment arms were pooled.
8. Reporting bias - Justification	Analysis reporting justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
9. Other bias - Assessment	Other risks of bias Is the study free from other sources of bias?	1= Yes, 4 = No		
9. Other bias - Justification	Other bias justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages). For	

			example, information is collected using a different survey instrument in different intervention groups; measurement of the intervention received in unclear.	
10. Blinding - observers - Assessment	Blinding of participants?	1=Yes 2=No 8=unclear 9= N/A	If there is no information, code NO. If there is information but it is ambiguous, code UNCLEAR.	
10. Blinding - observers - Assessment	Blinding of outcome assessors?	1=Yes 2=No 8=unclear 9= N/A	If there is no information, code NO. If there is information but it is ambiguous, code UNCLEAR.	
10. Blinding - analysts - Assessment	Blinding of data analysts?	1=Yes 2=No 8=unclear 9= N/A	If there is no information, code NO. If there is information but it is ambiguous, code UNCLEAR.	
10. Blinding - method(s)	Method(s) used to blind	Open answer (including describe method of placebo control)No 9= N/A	Describe method(s) used to blind	
11. External validity - Assessment	External validity	Open answer	a) What do authors say about external validity?	Include all information that can help assess the external validity of the results.

Provisional risk of bias assessment tool (QED)

General	ID	EPPI ID		
General	Time taken to complete assessment	Minutes		
General	Study first author	Open answer		
General	Outcome	Open answer		

General	Study design: What type of study design is used?	1= Natural experiment: randomised or as-if randomised 2= Natural experiment: regression discontinuity (RD) 3= CBA (non-randomised assignment with treatment and contemporaneous comparison group, baseline and end line data collection) – individual repeated measurement 4= CBA pseudo panel (repeated measurement for groups but different individuals) 5= Interrupted time series (with or without contemporaneous control group) 6= Panel data, but no baseline (pre-test) 7 = Comparison group with end line data only		
General	Methods used for analysis: Which methods are used to control for selection bias and confounding?	1 = Statistical matching (PSM, CEM, covariate matching) 2 = Difference in differences (DID) estimation methods 3 = IV-regression (2-stage least squares or bivariate probit) 4 = Heckman selection model 5 = Fixed effects regression 6 = Covariate adjusted estimation 7 = Propensity weighted regression 8 = Comparison of means	-	

		9 = Other (please state)		
General	Study population	Open answer	Provide any details in the paper that describe how the study population was selected, covering: a) How is the population selected? what is the sampling strategy to recruit participants from that population into the study? b) What are the characteristics of that study participants? c) Was this a pilot program aimed at being scaled up? d) Were there specific factors of success or failure in the implementation?	
General	Ethical clearance	Open answer	Provide any details of ethical research clearances granted. Report unclear if this information is not available.	
General	Study registration	Open answer	Provide any details of study registration, including registry IDs, etc.	

1: Selection bias - Assessment	1 - Mechanism of assignment: was the allocation or identification mechanism able to control for selection bias?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear		
1: Selection bias - Justification	For regression discontinuity designs	Open answer	a) Allocation is made based on a pre-determined discontinuity on a continuous variable (regression discontinuity design) and blinded to participants or; b) if not blinded, individuals reasonably cannot affect the assignment variable in response to knowledge of the participation decision rule; c) and the sample size immediately at both sides of the cut-off point is sufficiently large to equate groups on average.	Score “Yes” if criteria a), b), c) are all satisfied Score "Probably Yes" if there are minor differences in between both sides of the cut-off point but authors convincingly argue that the differences are unlikely to affect the outcome, OR individuals are not blinded and there are low risk of them affecting the assignment but the authors do not mention it. Score “Unclear” if it is unclear whether participants can affect it in response to knowledge of the allocation mechanism. Score "Probably No" if there are differences between individuals on both sides of the cut-off point, and there are doubts that the differences are due to individuals altering the assignment OR the participants are blinded but there is evidence that the decisions that determined the discontinuity is based on differences between the two groups or differences in time. Score “No” if the sample size is not

				sufficient OR there is evidence that participants altered the assignment variable prior to assignment. If the research has serious concerns with the validity of the assignment process or the group equivalence completely fails, we recommend assessing risk of bias of the study using the relevant questions for the appropriate methods of analysis (cross-sectional regressions, difference-in-difference, etc.) rather than the RDDs questions.
1: Selection bias - Justification	For assignment based non-randomised programme placement and self-selection (studies using a matching strategy or regression analysis, excluding IV)	Open answer	a) Participants and non-participants are either matched based on all relevant characteristics explaining participation and outcomes, or; b) all relevant characteristics are accounted for.** c) and the data set used contains relevant variable that are measured in a relevant way (i.e. they were not collected for a different purpose initially and therefore are good proxy for some characteristics). **Accounting for and matching on all relevant characteristics is usually only feasible when the programme allocation rule is known and there are no errors of targeting. It is unlikely that studies not based on randomisation or regression discontinuity can score “YES” on this criterion. There are different ways in which	Score “Yes” if a) or b) and c) are satisfied Score "Probably yes" if a) or b) are addressed for but there is some doubt related to c), OR authors combined statistical matching and difference-in-difference to cope with unobservable differences, OR they only did statistical matching and there was clear rules for selection into the program (no self-selection). Score “Unclear” if · it is not clear whether all relevant characteristics (only relevant time varying characteristics in the case of panel data regressions) are controlled. Score "Probably no" if only a statistical matching was done and there was self-selection into the program. Score “No” if relevant characteristics are

			covariates can be taken into account. Differences across groups in observable characteristics can be taken into account as covariates in the framework of a regression analysis or can be assessed by testing equality of means between groups. Differences in unobservable characteristics can be taken into account through the use of instrumental variables (see also question 1.d) or proxy variables in the framework of a regression analysis, or using a fixed effects or difference-in-differences model if the only characteristics which are unobserved are time-invariant	omitted from the analysis.
1: Selection bias - Justification	For identification based on an instrumental variable (IV estimation)	Open answer	Score “Yes” if an appropriate instrumental variable is used which is exogenously generated: for example, due to a ‘natural’ experiment or random allocation. Score "Probably yes" if there is less evidence (no balance table showing differences between the intervention and comparison group). Score “Unclear” if the exogeneity of the instrument is unclear (both externally as well as why the variable should not enter by itself in the outcome equation). Score "Probably no" if there is evidence that enrolment in the program is correlated with a variable that might also have an effect on outcome and on the instrumental variable.	

			Score “No” if it is clear that the instrument is not exogenous and affect the outcome through other channels than the program.	
2: Confounding - Assessment	2 - Group equivalence: was the method of analysis executed adequately to ensure comparability of groups throughout the study and prevent confounding?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear		
2: Confounding - Justification	For regression discontinuity design	Open answer	a) The interval for selection of treatment and control group is reasonably small OR authors have weighted the matches on their distance to the cut-off point; b) and the mean of the covariates of the individuals immediately at both sides of the cut-off point (selected sample of participants and non-participants) are overall not statistically different based on t-test or ANOVA for equality of means; c) Significant differences in covariates of the individuals have been controlled in multivariate analysis; and for cluster-assignment, authors control for external cluster-level factors that might confound the impact of the programme.	Score "Yes, if criterion a), b), c) and d) are addressed. Score "Probably yes" if b) is not addressed but c) is addressed and differences in means are not large. Score “Unclear” if insufficient details are provided on controls; or if insufficient details are provided on cluster controls. Score "Probably no" if b) is not addressed (absence of a difference test or balance table) and there are doubt regarding the continuity on both sides of the cut-off point (a). Score “No” otherwise.

2: Confounding - Justification	For non-randomised trials using difference-in-differences methods of analysis	Open answer	a) The authors use a difference-in-differences (or fixed effects) multivariate estimation method; b) the authors control for a comprehensive set of individual time-varying characteristics, and for cluster-assignment, authors control for external cluster-level factors that might confound the impact of the programme**; c) and the attrition rate is sufficiently low and similar in treatment and control, or the study assesses that drop-outs are random draws from the sample (for example, by examining correlation with determinants of outcomes, in both treatment and comparison groups); **Knowing allocation rules for the programme – or even whether the non-participants were individuals that refused to participate in the programme, as opposed to individuals that were not given the opportunity to participate in the programme – can help in the assessment of whether the covariates accounted for in the regression capture all the relevant characteristics that explain differences between treatment and comparison	Score "Yes, if a, b, c, d (if relevant) are addressed and baseline imbalances between groups were relatively low OR the method was combined by a statistical matching. Score "Probably yes" if all possible variables are controlled for and the selection into the program was done according to clear rules, but baseline imbalances between groups were very large. Score “Unclear” if insufficient details are provided; or if insufficient details are provided on cluster controls. Score "Probably no" if some time-varying characteristics are not controlled for and the program was self-selected by the intervention groups. Score “No” if any of the criterion is not addressed.
2: Confounding - Justification	For statistical matching studies including propensity scores (PSM) and covariate matching** **Matching strategies are	Open answer	a) Matching is either on baseline characteristics or time-invariant characteristics which cannot be affected by participation in the programme; and the variables used to match are relevant (for example, demographic and socio-economic	Score "Yes, if a, b, c, and d (if relevant) are addressed. Score "Probably yes" if the selection into the program was done according to clear rules, which are used for the matching but

	sometimes complemented with difference-in-difference regression estimation methods. This combination approach is superior since it only uses in the estimation the common support region of the sample size, reducing the likelihood of existence of time-variant unobservable differences across groups affecting outcome of interest and removing biases arising from time-invariant unobservable characteristics.		factors) to explain both participation and the outcome (so that there can be no evident differences across groups in variables that might explain outcomes); and, for cluster-assignment, authors control for external cluster-level factors that might confound the impact of the programme b) in addition, for PSM Rosenbaum's test suggests the results are not sensitive to the existence of hidden bias; c) and, with the exception of Kernel matching, the means of the individual covariates are equated for treatment and comparison groups after matching; d) different matching methods including varying sample sizes yields the same results and authors take into account the use of control observations multiple times against the same treatment in their standard error calculation.	there are slight imbalances remaining after matching. Score "Unclear" if relevant variables are not included in the matching equation, or if matching is based on characteristics collected at end line; or if insufficient details are provided on cluster controls. Score "Probably no" if the program was self-selected by the intervention groups or participants OR if the selection into the program was done according to clear rules but there is no baseline data available to match the participants or groups on. Score "No" if matching was done based on variables that are likely to be affected by the program or any other scenario that affect a), b) c) or d).
2: Confounding - Justification	For regression-based studies using cross sectional data (excluding IV)	Open answer	a) The study controls for relevant confounders that may be correlated with both participation and explain outcomes (for example, demographic and socio-economic factors at individual and community level) using multivariate methods with appropriate proxies for unobservable covariates, and, for cluster-assignment, authors control particularly for external cluster-level factors that might confound the impact of the programme; b) and a Hausman test with an appropriate instrument suggests there is no evidence of endogeneity**;	Score "Yes, if a, b, c and d are addressed. Score "Probably yes" if all criterion are addressed but authors did not report the Hausman test (b). Score "Unclear" if relevant confounders are controlled but appropriate proxy variables or statistical tests are not reported; or if insufficient details are provided on cluster controls. Score "Probably no" if any of the criterion other than b) is not addressed.

			c) and none of the covariate controls can be affected by participation; d) and either, only those observations in the region of common support for participants and non-participants in terms of covariates are used, or the distributions of covariates are balanced for the entire sample population across groups; **The Hausman test explores endogeneity in the framework of regression by comparing whether the OLS and the IV approaches yield significantly different estimations. However, it plays a different role in the different methods of analysis. While in the OLS regression framework the Hausman test mainly explores endogeneity and therefore is related with the validity of the method, in IV approaches it explores whether the author has chosen the best available strategy for addressing causal attribution (since in the absence of endogeneity OLS yields more precise estimators) and therefore is more related with analysis reporting bias.	Score “No" if none of the criterion are addressed.

2: Confounding - Justification	For identification based on an instrumental variable (IV estimation)	Open answer	a) The instrumenting equation is significant at the level of F≥10 (or if an F test is not reported, the authors report and assess whether the R-squared (goodness of fit) of the participation equation is sufficient for appropriate identification); b) the identifying instruments are individually significant (p≤0.01); for Heckman models, the identifiers are reported and significant (p≤0.05); c) where at least two instruments are used, the authors report on an over-identifying test (p≤0.05 is required to reject the null hypothesis); and none of the covariate controls can be affected by participation and the study convincingly assesses qualitatively why the instrument only affects the outcome via participation. If the instrument is the random assignment of the treatment, the reviewer should also assess the quality and success of the randomisation procedure in part a). d) and, for cluster-assignment, authors particularly control for external cluster-level factors that might confound the impact of the programme (for example, weather, infrastructure, community fixed effects, and so forth) through multivariate analysis.	Score "Yes, if a, b, c, d (if relevant) are addressed. Score "Probably yes" if one of the test required for criterion a) or b) is not reported but the other is, and the rest of the criterion are addressed and the instrument is convincing. Score “UNCLEAR” if relevant confounders are controlled for but appropriate statistical tests are not reported; or if insufficient details are provided on cluster controls Score "Probably no" if exogeneity of the instrument is not convincing and appropriate tests are not reported. Score “No” otherwise if any of the tests required for criterion a), b) or c) are reported and not satisfied.

3: Performance bias - Assessment	3 - Performance bias: was the process of being observed free from motivation bias?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) For data collected in the context of a particular intervention trial (randomised or non-randomised assignment), the authors state explicitly that the process of monitoring the intervention and outcome measurement is blinded, or argue convincingly why it is not likely that being monitored could affect the performance of participants in treatment and comparison groups in different ways (such as resulting in Hawthorne or John Henry effects). b) The study is based on data collected in the context of a survey, and not associated with a particular intervention trial, or data are collected from administrative records or in the context of a retrospective (ex post) evaluation.	Score “Yes” if either criterion a) or b) are satisfied; Score "Probably yes" if the study is based on survey data collected during a trial and there is no obvious issue with the monitoring processes but authors do not mention potential risks. Score “Unclear” if it is not clear whether the authors use an appropriate method to prevent Hawthorne and John Henry Effects (e.g. blinding of outcomes and, or enumerators, other methods to ensure consistent monitoring across groups). Hawthorne effects may result where participants know that they are being observed and John Henry Effects may result from participant knowledge of being compared. Score "Probably no" if there was imbalance in the frequency of monitoring in intervention groups, which might have influenced participants' behaviours. Score "No"
3: Performance bias - Justification	Performance bias - Justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	

4: Spill-overs, cross-overs and contamination - Assessment	4 - Spill-overs, cross-overs and contamination: was the study adequately protected against spill-overs, cross-overs and contamination?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) There was no implementation issues that might have led the control participants to receive the treatment (implementer's mistake). b) The intervention is unlikely to spill-over to comparisons (e.g. participants and non-participants are geographically and/or socially separated from one another and general equilibrium effects are not likely) or the potential effects of spill overs were measured (e.g. variation in the % of unit within a cluster receiving the treatment). c) There is no risk of contamination by external programs: the treatment and comparisons are isolated from other interventions which might explain changes in outcomes. d) There is nothing in the surveys that might have given the control participants an idea of what the other group might receive OR they did but there is no risk that this has changed their behaviours; AND the survey process did not reveal information to the control group that they did not have before (e.g. the study aims to measure increase in take up of a service or product that participants might not know about) Authors might put something in place in the design of the study that allows to control for that survey effect (e.g. a pure control with no monitoring except baseline end line)	Score “Yes” if criterion a), b), c) and d) are satisfied; Score "Probably yes" if there is no obvious problem but there is no information reported on potential risks related to spill overs, contamination, or survey effects in the control group OR if there were issues with spill-overs but they were controlled for or measured. Score “Unclear” if spill-overs, cross-overs, survey effects and/or contamination are not addressed clearly. Score "Probably no" if any of the criterion a), b), c) or d) are not satisfied but the scale of the issue is not clear. Score “No” if any of the criterion a), b), c) or d) are not satisfied and happened at a large scale in the study.

4: Spill-overs, cross-overs and contamination - Justification	Spill-overs, cross-overs and contamination - Justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
5: Outcome measurement bias - Assessment	5 - Outcome measurement bias	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) Outcome assessors are blinded or the outcome measures are not likely to be biased by their judgement. b) For self-reported outcomes: respondents in the intervention group are not more likely to have accurate answers due to recall bias; c) For self-reported outcomes: respondents do not have incentives to over/under report something related to their performance or actions, OR researchers put in place mechanisms to reduce the risk of reporting bias (researchers not strongly involved in the implementation of the program and it is clear that their answers to the survey will not affect what they receive in the future) OR authors have measured the risks of bias through falsification tests or measuring the effect on placebo outcomes in cases where there was a risk of reporting bias. d) Timing issue: the data collection period did not differ between intervention and comparison group, the baseline data is not likely to be affected by the beginning of the intervention or affects a small percentage of the study participants.	Score “Yes” if criterion a), b), c) and d) are satisfied: Score "Probably yes" if there is a small risk related to any of a), b), c) or d) and there is no more information provided to justify the absence of bias OR if there was a high risk of bias but authors have either controlled it in their design or measured it with a placebo outcomes. Score “Unclear” if it there is a high risk related to any of a), b), c) or d) and there is no more information provided to justify the absence of bias. Score "Probably no" if there are high risk related to a), b), c) or d) and it is clear that authors were not able to control for this bias. Score “No” if there is evidence of bias.

5: Outcome measurement bias - Justification	Outcome measurement bias - Justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
6: Reporting bias - Assessment	6 - Selective analysis reporting: was the study free from selective analysis reporting?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) a pre-analysis plan is published, especially for prospective NRS but it should also be for retrospective studies b) authors use 'common' methods of estimation (i.e. credible analysis method to deal with attribution given the data available) ; c) There is no evidence that outcomes were selectively reported (e.g. results for all relevant outcomes in the methods section are reported in the results section) ; d) Requirements for specific methods of analysis: - For PSM and covariate matching: (a) Where over 10% of participants fail to be matched, sensitivity analysis is used to re-estimate results using different matching methods (Kernel Matching techniques); (b) For matching with replacement, no single observation in the control group is matched with a large number of observations in the treatment group. - For IV (including Heckman) models, (a) The authors test and report the results of a Hausman test for exogeneity ($p \leq 0.05$ is required to reject the null hypothesis of exogeneity); (b) the coefficient of the selectivity correction term (Rho) is significantly different from zero ($P<0.05$) (Heckman approach).	Score "Yes" if a), b), c) and d) are satisfied OR if a) is not met and it is a retrospective NRS. Score "Probably Yes" if authors combined methods and reported relevant tests (d) only for one method OR if all the criteria are met except for a) and it is a prospective NRS Score "Unclear" if intended outcomes not specified in the paper OR if any of the requirements for d) are not reported. Score "Probably No" if b) is addressed, but authors did not present results for all outcomes announced in the method section OR did not meet requirement d) although reported. Score "No" if authors use uncommon or less rigorous estimation methods such as failure to conduct multivariate analysis for outcomes equations OR if some important outcomes are subsequently omitted from the results or the significance and magnitude of important outcomes was not assessed.

			- For studies using multivariate regression analysis, authors conduct appropriate specification tests (e.g. testing robustness of results to the inclusion of additional variables, or (very rare) reporting results of multicollinearity test etc).	
6: Reporting bias - Justification	Analysis reporting bias - Justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
7: Other bias - Assessment	7 - Other risks of bias: Is the study free from other sources of bias?	1= Yes, 4 = No	Score “Yes” if the reported results do not suggest any other sources of bias. Score “No” if other potential threats to validity are present, and note these here (e.g. coherence of results, survey instruments used are not reported)	
7: Other bias - Justification	Other risks of bias - Justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
8: External validity	8 - External validity	Open answer	Open answer- what do authors say about external validity, if anything?	