Online appendix F

Quality assessment of impact evaluations

1. *Study design (potential confounders taken into account):* impact evaluations need either a well-designed control group (preferably based on random assignment) or an estimation technique which controls for confounding and the associated possibility of selection bias.
2. *Power calculations:* Small sample size can result in an under-powered study with a high risk of not detecting an effect from the intervention when there actually is one. The combination of under-powered studies and publication bias can put an upward bias on the assessment of the overall effect size from a body of literature. The problem of sample size is addressed by conducting power calculations before the study to determine the required sample size. We will not use this item in the overall assessment of the study. However coding mention of power calculations signals the importance of both conducting and reporting power calculations.
3. *Attrition or losses to follow up*: can be a major source of bias in studies, especially if there is differential attrition between the treatment and comparison group (called the control group in the case of RCTs) so that the two may no longer be balanced in pre-intervention characteristics. The US Institute of Education Sciences What Works Clearing House (WWC) has developed standards for acceptable levels of attrition, in aggregate and the differential, which are applied here.
4. *Description of intervention:* If the intervention is not well described then the evidence may be misinterpreted to support an intervention not actually supported by study findings. We rate as low if the description is just named, medium if there is a short description, and high if there is a detailed description. We do not use this item in the overall assessment of the study.
5. *Definition of outcomes:* Outcomes should be clearly defined so that study findings can be properly interpreted. So far as possible, unless a subjection perception is required, that questions should rely on objective factors, and utilize data collection instruments

which have been validated for the context in which they are being applied. We rate as high of there is clear definition of the outcome and how it is being measured, or reference to an existing tool. Medium rating is given is if there is a brief description, and low if the outcome is named but not adequately described.

6. *Baseline balance:* Baseline balance means that the treatment and comparison groups have the same average characteristics at baseline, not only for outcomes but other factors which may affect the impact of the programme such as a prior history of parental alcohol abuse. We rate low confidence on study findings if baseline balance is not reported for non-RCTs or it is reported and there is a significant difference of 10 or more than 10 percent, medium confidence if imbalance is between 5-10 percent, and high if an RCT or if imbalance is 5 or less than 5 percent.
7. *Representativeness of a large-scale intervention: for a study to be useful the study results population must be clearly defined. A) Is the sampling frame clearly defined? B) Does the sampling frame include at least 1000 beneficiaries or covers an administrative area larger than a village? C) Is the sample randomly drawn from this sampling frame? We rate as high if all questions are answered with a "yes". We rate as medium if at least question B) is answered with a "yes" and C) is <u>not</u> answered with a "no". Otherwise, we rate as low.*
8. *Precision of regression estimates: Does the study mention clustering in the design? Are the standard errors clustered at the level of clustering (likely the level of intervention randomisation). We rate the study from high, medium to low, based on the clustering of standard errors at the level of clustering, at a higher level, or without any clustering at all, respectively.*

Overall assessment: The overall assessment uses a weakest link in the chain principle so that the overall assessment is the lowest of assessment given to any of the relevant items. As mentioned above, not all items are used in this assessment. So the overall assessment is the lowest of the assessments for items 1a, 4, 6, 7 and 8.

Table 1. Quality assessment procedure

	ITEM	POINT IN TIME (WHERE APPLICABLE)	RATING
1A	Study design (Potential confounders taken into account)		High confidence: RCT, RDD, instrumental variable under LATE Medium confidence: DiD with matching, PSM Low confidence: other matching, DiD alone, instrumental variable otherwise

1B	Masking or blinding (RCTs only)	DO NOT CODE	High confidence: any blinding or any mention of blinding Medium confidence: no blinding Low confidence is not used for this item
2	Power calculations are reported	DO NOT CODE	High confidence: any mention of power calculations as basis for sample size Medium confidence: no mention of power calculations Low confidence is not used for this item
3	Losses to follow up are presented and acceptable	DO NOT CODE	High: attrition within IES bounds Medium: attrition close to IES bounds Low: attrition not reported or attrition outside IES bounds N/A for ex post studies
4	Intervention is clearly defined		High confidence: intervention clearly and fully described Medium confidence: brief description of intervention Low confidence: intervention named but not described, or not named
5	Outcome measures are clearly defined and reliable		High confidence: outcome measure clearly and fully described, preferably with reference to validation Medium confidence: brief description of outcome Low confidence: outcome named but not described
6	Baseline balance (N.A. for before versus after)		High confidence: RCT or baseline balance report and satisfactory (imbalance on 5 or less than 5 percent) Medium confidence: Imbalance in 5-10 percent baseline variables Low confidence: Baseline balance not reported, or reported and lack of balance on 10 or more than 10 percent of baseline variables

7	Representativeness of a large-scale intervention	*A) Is the sampling frame clearly defined?* *B) Does the sampling frame include at least 1000 beneficiaries or covers an administrative area larger than a village?* *C) Is the sample randomly drawn from this sampling frame?* High confidence: all three answered as yes Medium confidence: *at least question B) is answered with a "yes" and C) is <u>not</u> answered with a "no* Low confidence: None of the above
8	Precision of estimate (in case of regressions)	High confidence: standard errors are clustered by intervention design level Medium confidence: standard errors are clustered but not clear at what level Low confidence: standard errors are not clustered Not applicable for studies where no clusters are mentioned.
	Overall confidence in study findings	Low confidence: Low rating across any of the items 1a, 4, 5, 6, 7 and 8 Medium confidence: Medium rating across any of the items 1a, 4, 5, 6 7 and 8 (and no LOW rating) High confidence: High rating across all of the items 1a, 4, 5, 6, 7 and 8