THE CAMPBELL COLLABORATION

Protocol:
Farmer Field Schools for Improving Farming Practices and Farmer Outcomes in Low- and Middle-income Countries: A Systematic Review

Hugh Waddington, Birte Snilstveit, Jorge Garcia Hombrados, Martina Vojtkova, Jock Anderson, and Howard White

Submitted to the Coordinating Group of:

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

Plans to co-register:

☐	No		
☒	Yes	☐ Cochrane	☐ Other
☐	Maybe		

Date Submitted: 26 January 2012
Date Revision Submitted: 27 July 2012
Approval Date: 1 September 2012
Publication Date: 01 November 2012

BACKGROUND

Description of the condition

Agriculture has wide-ranging global impacts which extend to economic growth, poverty reduction, food security, livelihoods, rural development and the environment (Green et al., 2005). Agriculture is the main source of income for around 2.5 billion people in the developing world (FAO, 2003, p. 1). In addition, around 70 percent of the global extreme poor – or over one billion people – lives in rural areas in low and middle income countries (IFAD, 2010, p.233), most of whom rely directly or indirectly on agriculture for their livelihoods. Investment in agriculture has been shown to have beneficial impacts on agricultural growth and poverty reduction (Fan & Rao, 2003). Moreover, the poorest population quintiles benefit significantly more from agricultural growth than growth in other sectors of the economy (United Nations, 2008; World Bank, 2007).

The modernisation of farming practices in the 1960s and 70s during the 'Green Revolution' improved agricultural yields substantially in those areas it reached and raised national production and food security (IFAD, 2001). However, two key challenges emerged (van den Berg & Jiggins, 2007). The first problem was that poor farmers were being left behind, particularly in sub-Saharan Africa where many were not reached by modernisation approaches. In addition, those technologies promoted were not appropriate to the challenges facing smallholders in the African context, particularly women farmers (Inter-Academy Council, 2004). Second, modernisation was also associated with adverse environmental and health consequences, relating to water pollution, declining soil quality, soil erosion, pest resistance and loss of biodiversity.

A particular problem emerged around environmental and health consequences of chemical pesticides use. Chemical pesticides have been heavily promoted and publicly subsidised under the modernisation agenda to such an extent that their overuse led to insect pests becoming resistant and causing major outbreaks of insect pests in rice crops in Asia in the 1970s and 80s. In addition, prolonged exposure to pesticides was associated with chronic and acute health problems among rural residents (Pingali & Roger, 1995). Use of broad-spectrum insecticides in agriculture has even been linked to mosquito vectors of malaria developing resistance to insecticides used in malaria control programs (Diabate et al., 2002; cited in van den Berg & Jiggins, 2007).

Description of the intervention[1]

Ensuring that farmers have appropriate knowledge and skills to deal with changing

[1] This study is part of a larger systematic review on agricultural extension services (Waddington et al., 2010).

environments, and that they have access to technology and management practices appropriate to their needs, are key components of sustainable agricultural development strategies. Agricultural extension and advisory services (hereafter extension services) are used by policy makers to achieve this aim, and comprise "the entire set of organisations that support and facilitate people engaged in agricultural production to solve problems and to obtain information, skills and technologies to improve their livelihoods" (Anderson, 2007, p.6). Extension was traditionally viewed as a means of transferring technologies developed in research stations and farm management practices to farmers, and used 'top-down' institutions of delivery, as characterised for example by the World Bank's Training and Visit System (Gautam & Anderson, 2000).

These traditional extension approaches were criticised for providing a 'one size fits all' approach (Birner et al., 2006) which failed to factor in the diverse socio-economic and institutional environments faced by farmers, for failing to involve farmers in the development of technology and practices appropriate to their contexts, or to empower them more generally as problem solving decision makers. Ultimately extension has failed to achieve its main objective of farm productivity improvements, particularly in Africa (Anderson, 2007; Birkhaeuser et al., 1991). In addition, more intensive approaches were considered necessary to disseminate complex messages, such as on sustainable pest management. Since the 1980s, the approach to extension service delivery has drawn increasingly on more participatory methods, of which farmer field schools have become prominent (van den Berg & Jiggins, 2007).[2] Participatory approaches to extension are based on the idea that they create spaces for farmer 'self-learning' and sharing and also allow the agents and agricultural researchers to learn from the farmers (Birner et al., 2006).

Farmer field schools (FFSs) originated in Asia as a means of achieving several objectives, of which key was to deliver training on 'integrated pest management' as an alternative to intensive pesticide spraying, which was severely damaging farm production, the environment and farmers' health. Integrated pest management (IPM) was developed in the 1960s and 70s (Kelly, 2005) and aimed to minimise pesticide use through use of more 'natural' techniques of pest management. Integrated pest management methods promoted in FFSs typically range from more simple practices, such as not applying pesticides in the first 30 days after planning ('no early spray'), to more complex ones that require in-depth agro-ecological and crop management knowledge, such as being able to differentiate beneficial from harmful insects, and creating a conducive environment for pest predators (Ricker-Gilbert et al., 2008).

FFSs are participatory, learner-centred and characterized by heavy reliance on learning-by-

[2] There has been a similar evolution in the use of more bottom-up approaches to technology development through agricultural research, such as the local agricultural research committees (CIALs) approach (Braun et al., 2000).

doing (Pontius et al., 2002). Farmer field schools use intensive 'discovery learning' techniques to provide farmers with the skills and confidence to adopt different growing techniques and change the mix of inputs used on their farms. In the case of IPM-FFSs, field school participants are instructed on how to move away from pesticides to more natural techniques of pest management. Objectives of the schools include increasing farm productivity, reducing negative environmental impacts and promoting farmer empowerment.

The FFS approach draws on participatory methods, both in terms of its bottom-up focus, with curricula drawing on priorities identified by farmers, and in terms of the focus on farmer experimentation and building problem solving capabilities, which empowers farmers "to handle their own on-farm decisions, using experiential learning techniques developed for non-formal adult education purposes" (Khisa, 2004). Thus, FFSs aim to provide farmers with skills which enable them to solve problems for themselves and the group activities empower farmers both within their own communities and outside. Indeed, the approach is far more intensive than other extension delivery mechanisms and broader in its objective to provide skills and empower individuals, and is therefore considered an adult education intervention by some practitioners and analysts (Braun et al., 2006; van den Berg & Jiggins, 2007).[3]

Originally developed for rice crops in Indonesia in the 1980s by the United Nations Food and Agriculture Organisation (FAO), FFSs had been implemented in 87 countries worldwide and produced 10-20 million field school graduates by 2008 (Braun & Duveskog, 2008). While all FFSs should be based on the same process, the approach can be adopted to suit particular needs, crops or contexts (Pontius et al., 2002). Thus, as FFSs have been promoted around the world, the IPM curriculum has been modified depending on the context, and applied to other food staples, vegetables and cotton (Braun & Duveskog, 2008). In Africa, Integrated Production and Pest Management (IPPM) has been promoted, which reflects a more 'holistic' approach to improving production, in which pests and pesticide use are not necessarily the main production problems (Stathers et al., 2005).[4] Other variants include integrated disease management (IDM), integrated crop management (ICM), integrated plant nutrient management (IPNM), and integrated water and soil management (IWSM)[5] In

[3] This distinction is not merely academic, but central to discussions about cost-effectiveness.

[4] Drawing on the lessons of IPM, Integrated vector management (IVM) is being applied in the health sector to combat malaria and other vector-borne diseases http://www.who.int/malaria/vector_control/ivm/en/index.html. This variant is beyond the scope of this review.

[5] http://www.fao.org/agriculture/crops/core-themes/theme/spi/scpi-home/managing-ecosystems/integrated-plant-nutrient-management/en/

addition, the farmer field school curriculum has also been broadened to tackle populations in particular contexts, such as Junior Farmer Field and Life Schools (JFFLS) which have been implemented among youths across Africa and include HIV-risk reduction in addition to agriculture components more standard to FFSs (Braun & Duveskog, 2008).

How the intervention might work

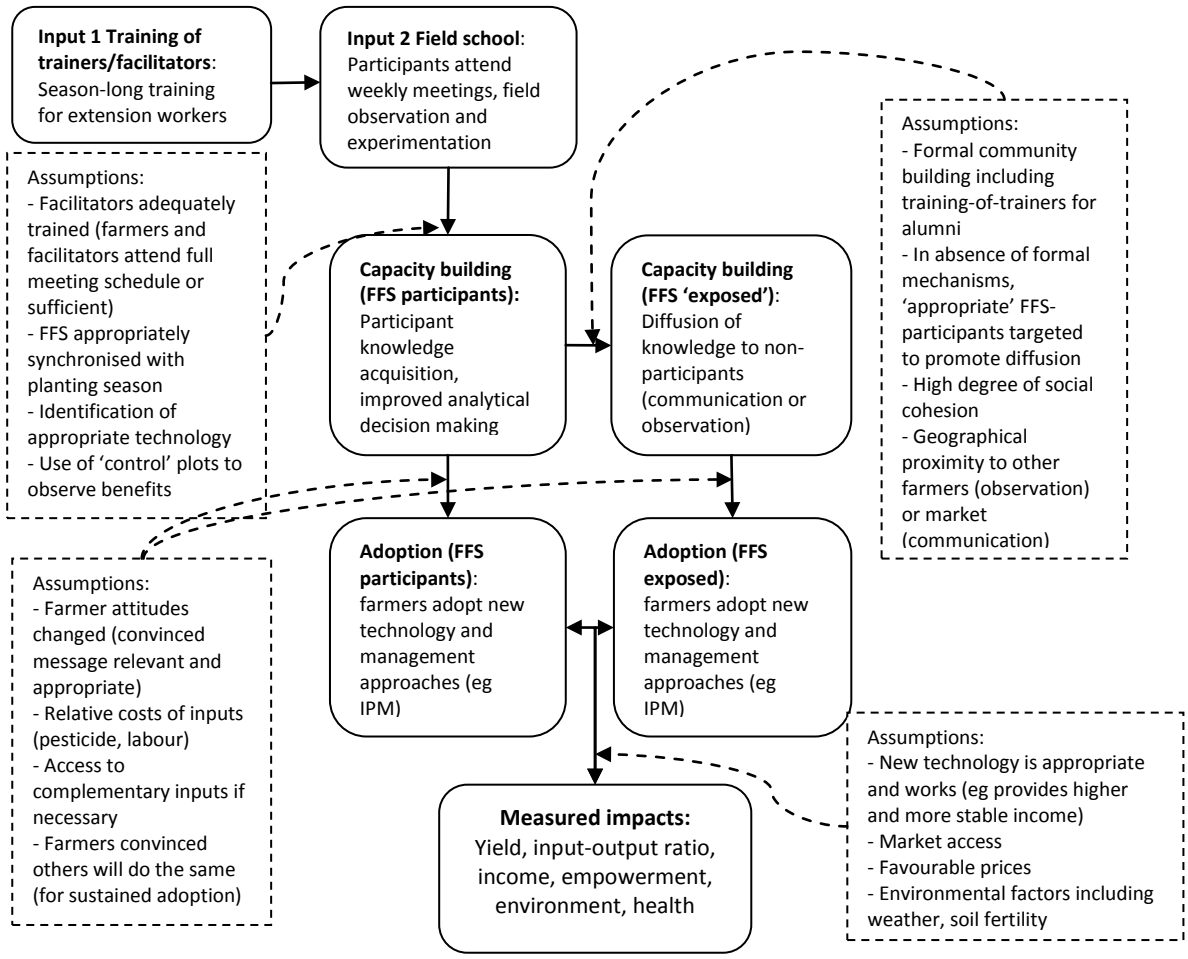
A typical FFS includes a field-based season-long training program delivered by a FFS facilitator, with weekly meetings nearby the plots of participating farmers (Pontius et al., 2002). Each FFS typically has from 25 to 30 participants, with farmers working together in groups of five. Facilitators can be either extension agents or selected graduates from previous FFSs, who undergo a training-of-trainers course tailored to equip them to facilitate field schools (Braun & Duveskog, 2008). The facilitators use experiential, participatory and learner-centred educational methods, including use of demonstration plots comparing farmers' existing practices with new practices promoted in the FFS to enable farmers to observe benefits (Pontius et al., 2002). Field school activities include agro-ecosystem analysis, farmer experimentation, activities to improve group dynamics, field days focusing on specific local problems and participant presentation of course material and the results of their studies (ibid).

Due to the externalities associated with pesticides use (social costs exceed private costs), it may be important that integrated pest management practices are adopted throughout a community for the approach to be sustainable, particularly in extreme situations of pesticide misuse. As Feder et al. (2004a) note, lack of adoption of IPM practices by neighbouring farmers might curtail the effectiveness of the intervention, as pests from their fields may re-infest the fields of adopters, eventually leading to disadoption of IPM by FFS participants. To promote community-wide uptake of IPM, FFS participants are encouraged to share their knowledge with non-participants as a way of promoting farmer-to-farmer diffusion (Feder et al., 2004b). The extent to which it is assumed this will occur informally through FFS graduates' social networks or needs to be encouraged through community initiatives including formal training-of-trainers programmes for alumni, seems to vary from programme to programme. However, FAO guidelines on 'community-IPM' which focuses on institutionalising IPM at the local level, indicate formal approaches involving FFS alumni are necessary: 'without post-FFS educational opportunities, there will be no community movement' (Pontius et al., 2002).

A wide range of factors is likely to influence effectiveness of farmer field schools. **Figure 1** presents a stylized causal chain linking farmer field school delivery inputs with final outcomes, via intermediate outcomes in terms of capacity building and technological adoption. Intermediate outcomes are shown for both field school participants and for 'exposed' farmers who benefit via farmer-to-farmer diffusion knowledge spillovers, and are expected to be those living in close geographical proximity to field school participants or involved in their social networks.

Underlying each link in the causal model are a number of assumptions which determine the extent of behaviour change and therefore the extent to which impacts materialize in practice. For instance, facilitators are a key input: they may be 'traditional' extension agents who have received training in the FFS approach, meaning they are required to move away from the top-down approaches to which they are familiar, and adopt a more participatory, learner-centred approach (Feder et al., 2004a). The extent to which they are able to do this may have implications for programme effectiveness. Similarly, the relevance of the FFS curriculum to farmers, to the extent that the new practices are actually appropriate, will influence farmers' attitudes and behaviour change. Farmers also need to be trained adequately, in that they have attended sufficient meetings over the planting season. Observability of improvements based on the comparison of experimental FFS with 'control' plots which use standard approaches to pest management, also appear to be a key component. Finally, the price of pesticide relative to (opportunity costs of) labour is also likely to determine adoption of IPM, where it involves substantial increases in demands on farmers' time.

Figure 1 FFS theory of change



Source: authors

Characteristics of the intervention are important moderators of both the causal chain framework presented here and the likely impacts. As noted in Davis et al. (2010), not all FFSs are the same. Earlier FFS programmes implemented in Asia usually focused on rice IPM, while later ones focused on different crops and livestock. However, the extent to which FFSs adapted to different contexts are based on the same process, both in the design and implementation of the programmes, is likely to vary. For instance, different FFSs use different approaches, with some following a transfer-of-technology approach, while others focus to a greater extent on education or empowerment (Davis et al., 2010). These are important contextual differences which influence the way in which the concept of FFS is operationalised on the ground and may therefore also have implications for intervention effectiveness.

Characteristics of local communities, such as heterogeneity in terms of land- and asset-holdings, ethnicity, education, gender roles and the degree of social cohesion, will determine the ability of the schools to reach appropriate beneficiaries, including disadvantaged farmers such as women. In the absence of formalised community building and training-of-trainers programmes for alumni, characteristics of local communities may be important determinants of the degree of diffusion of knowledge and practices from participants to non-participants. The assumption that there will be some diffusion between farmers may not be an unreasonable one in principle for simple practices. However, evidence also suggests that, in the case of technically complex issues or costly technologies, farmers prefer first-hand knowledge or advice from specialized information sources such as experts (Feder et al. 2004a). As FFSs are used to disseminate complex information and aim to improve participants' decision-making skills, FFS graduates may be limited in their ability to transmit all but the simplest of messages effectively to other farmers through informal means. Whether the diffusion mechanism is informal or formalised will therefore have implications for beneficiary targeting (Feder & Savastano, 2006). Without formal mechanisms, participants would ideally be selected if they have characteristics which will enhance diffusion, such as those respected in their communities and those with strong social networks. This may conflict with other objectives of FFS, such as targeting women farmers.

Finally, the effectiveness of the schools in fostering adoption of the new technology and improved agricultural outcomes will depend on contextual factors. External factors, notably weather conditions, soil fertility, plant disease and climate trends, determine production and yields. Market prices and market access, both to purchase inputs and sell produce, determine the value of production and therefore farmer income. It is also possible that the technologies promoted by FFSs do not act to change yields (the amount of crop produced per unit of land area), but still act to improve income and net revenues (value of production less input costs) by reducing pesticide costs, provided these are not offset by any net increases in costs of labour in applying the new technologies. Moreover, in contexts where pesticides and pest management are not necessarily the key constraints to production, improvements in productivity may not necessarily arise from reduced pesticides use but as a result of adoption

of the other practices being promoted, such as soil management.

Why it is important to do this review

Since the 1980s there has been a decline or stagnation in public expenditure on agriculture in most developing countries (Akroyd & Smith, 2007). Likewise, the proportion of official development assistance (ODA) going to agriculture is estimated to have declined from around 18 per cent in 1979 to 3.5 per cent in 2004 (World Bank, 2007, p. 41). However, as noted in the World Development Report on Agriculture, "extension services, after a period of neglect, are now back on the development agenda... [but] More evaluation, learning, and knowledge sharing are required to capitalize on this renewed momentum" (World Bank, 2007, p. 175). Poverty reduction strategies in 24 African countries also listed extension as a top agricultural priority (InterAcademy Council, 2004; cited in Davis, 2006). However, age old questions remain including how to raise yields and farmer incomes, how to do so in an environmentally sustainable manner, and how to bring extension services to the poorest people.

A large literature exists measuring the impacts of farmer field schools, as summarised in a number of literature reviews (Davis, 2006; Feder et al., 2010; Quizon et al., 2001; Tripp et al., 2005; van den Berg, 2004; van den Berg & Jiggins, 2007). The reviews provide conflicting conclusions about the effectiveness of farmer field schools. Further, none appears to draw on a systematic search for all available literature, applying standard inclusion criteria, and critically appraising and synthesizing literature. Moreover, most of them draw on studies that appear suspect in terms of causal validity, due to problems of confounding, selection bias and spillovers, and thus are liable to high risks of bias in attributing outcomes to the intervention.[6]

Van den Berg (2004) provides the most authoritative review of IPM-FFS to-date, synthesising 25 evaluation studies. Most studies focused on rice and measured immediate impact of the FFSs in terms of reduced pesticide use and changes in yields, reporting considerable reductions in pesticide use, with some studies also showing an increase in yields. The review concludes: "Studies reported substantial and consistent reductions in pesticide use attributable to the effect of training. In a number of cases, there was also a convincing increase in yield due to training.... Results demonstrated remarkable, widespread and lasting developmental impacts" (p.3). Building on the latter, Van den Berg and Jiggins (2007) argue that FFSs have had additional benefits to that of IPM including facilitating collective action, leadership, organisation and improved problem-solving skills. However, the methodology of the studies reviewed varies, and a large number are based on pretest-

[6] There are meta-analyses on the effects of agricultural extension more generally, notably Alston et al. (2000). However, these studies frequently draw on economic appraisals based on weak counterfactuals.

posttest evaluation design with no non-intervention comparison group. Moreover, the conclusions of these reviews are based on significance-based vote counting, rather than sample weighted meta-analysis of effects.

Tripp et al. (2005) note that despite the sizeable investments in FFS in Asia, there is a lack of rigorous evidence on the effectiveness of the approach. Their review includes seven studies on the effects of FFS on insecticide use in addition to their own study from Sri Lanka, and all but one of these studies report that FFS participants reduced their pesticide use. However, all of these studies suffer from methodological weaknesses, due to lack of non-FFS comparison groups and high risk of selection bias. The authors conclude that while "the FFS approach has undoubtedly succeeded in lowering insecticide use in a number of Asian rice examples, judgments on its overall impact await further study" (p. 1711). The authors also found little evidence to suggest effective diffusion of knowledge between FFS participants and non-participants, nor sufficient evidence to conclude that FFS groups continue on their own.

FFS is a particularly intensive intervention, with high costs in terms of both facilitation and opportunity costs of beneficiaries' time. Leading authors from the literature have therefore noted that FFS is unlikely to be a solution to extension delivery, and only likely to be scalable under certain circumstances, and to face particular challenges (Braun et al., 2006; Davis, 2006). Braun and Duveskog (2008) argue that relative cost-effectiveness of FFS should be put in the context of rural adult education rather than extension "when FFS are regarded as a form of public investment in farmer education to tackle rural poverty – and hence as a tool for achieving the Millennium Development Goals" (p.19). Van den Berg and Jiggins (2007) also note that discussions on the fiscal sustainability of FFSs should include considerations of who will pay for the externalities of pesticide use; they conclude that the evidence gathered in their review suggests that FFSs can be a sustainable way of increasing farmers' skills and thus contributing towards escaping poverty. However, Quizon et al. (2001) note that lack of fiscal sustainability is a generic problem affecting many large-scale public extension services, concluding that FFSs face the same issues as other approaches. The cost per farmer is likely to be high compared to less intensive extension approaches and the evidence from Indonesia suggests there is a low rate of informal diffusion from direct beneficiaries of the schools to non-beneficiaries based locally.[7] They suggest that as the situation for farmers, in terms of political power, governance systems and day-to-day interactions among farmers, is quite similar in many other developing countries in Asia and Africa, the results are relevant for discussions of similar extension activities in these areas. They warn that while pilot projects might indicate the viability of the FFS approach in certain

[7] In Bangladesh, Ricker-Gilbert et al. (2008) estimate FFS per capita costs, including opportunity costs of farmer and trainer time, at over ten times those of other extension approaches including demonstration field days and extension agent visits.

circumstances, the issue of fiscal sustainability is particularly relevant when scaling up.

The existing reviews provide some suggestive evidence of the effects of FFS, but come to widely different conclusions in a hotly debated, policy important field. Our systematic review aims to shed light on this debate by aiming to provide a systematic and exhaustive literature search, together with a comprehensive and unbiased synthesis of the existing evidence.

Our review will contribute to the discussion, and systematic review methods development more generally, by synthesising both quantitative and qualitative literatures. To ensure the review is adequately oriented towards both reporting effects and explaining the reasons for them, we will synthesise effects along the causal chain, together with qualitative evidence. For quantitative synthesis, where studies are judged sufficiently similar to do so, we will use meta-analysis to pool study effects. An important benefit of meta-analysis is that it aims to account for the statistical power of studies. The alternative to meta-analysis is 'vote counting', which simply reports the numbers of positive, negative and insignificant findings. However, using this approach, one is more likely to come to incorrect conclusions about studies that find insignificant effects on outcomes, when this could simply be the result of low statistical power (small sample size), and therefore about the body of literature as a whole (Borenstein et al., 2009). One of the main benefits of meta-analysis is that by pooling across a larger sample, it takes into account both magnitude and precision of effects, allowing the researcher to correct for possibly under-powered studies.

OBJECTIVES

The primary objective of the review is to synthesise evidence on the effectiveness of farmer field school interventions in disseminating information on integrated pest management (IPM). The review aims to answer the following questions:

Review question (1): What is the impact of farmer field schools on their objectives in terms of 'endpoint' outcomes such as increased yields, net revenues and farmer empowerment, and intermediate outcomes such as capacity building and adoption of improved practices (e.g. reduced use of pesticides) in low and middle income countries?

Review question (2): Under which circumstances and why: what are the facilitators and barriers to FFS effectiveness and sustainability?

METHODS

The review will follow Campbell and Cochrane Collaboration approaches to systematic reviewing (Becker et al., n.d.; Hammerstrøm et al., 2010; Higgins & Green, 2011; Shadish & Myers, 2004; Shemilt et al., 2008). The review is also informed by theory-based impact evaluation (White, 2009) using the theory of change (Figure 1) as the framework for the review, to guide the types of studies included, data collection and analysis. The review will

systematically collect and synthesize quantitative evidence from high quality impact evaluations of farmer field school interventions using meta-analysis, to answer *review question (1)*. Outcomes will be synthesised along the causal chain, from intermediate outcomes such as capacity building, technological adoption and diffusion to final outcomes such as agricultural yields, household income and other indicators of household wellbeing.

Farmer field schools are complex interventions implemented in a range of different contexts, making the limitations of a systematic review focusing solely on effectiveness particularly apparent. For the review to be more useful for policy-makers and practitioners, we will extend the review of effectiveness by including quantitative and qualitative studies to address *review question (2)*, focusing on underlying factors that determine or hinder the effectiveness of FFSs. We will conduct the two syntheses in parallel, before integrating them in a final synthesis. Our methodology is also informed by chapter 20 in the Cochrane Handbook (Higgins & Green, 2011), the additional guidance developed by the Cochrane Qualitative Methods Group (Noyes et al., 2011) and the increasing number of examples of systematic reviews in international development based on or incorporating qualitative evidence (e.g., Munro et al., 2007; Williamson et al., 2009).

Study selection criteria

Studies will be included in the review if they meet the following selection criteria.

Types of participants

The review includes arable farmers, living in developing (low- or middle-income) countries, as defined by the World Bank, at the time the intervention was carried out. For studies to be included, they need to collect and report on data at the farm or household level. Many of the included populations are by definition disadvantaged, but studies focusing on particular disadvantaged groups, or conducting analysis across disadvantaged groups, will be included in the review.

The review excludes livestock farmers, who receive different types of training than arable farmers, and farmers based in high-income countries where the challenges facing farmers in terms of poverty, land size, crops, and agro-ecological and environmental contexts are usually very different.

Types of interventions

Studies must report specific FFS-IPM interventions, or similar. Interventions are identified as farmer field schools that contain all the following components:

- Involved intensive, facilitated group training, normally involving season-long weekly meetings and use of ‘control’ plots, which were farmed using ‘standard’ practices.

- Field schools providing information on holistic approaches to inputs use, such as reducing use of pesticides and insecticides or improved use of fertiliser or other production practices and disease control methods, through 'integrated pest management', 'integrated production and pest management', 'integrated crop management', 'integrated disease management' and so on.

- Studies are also eligible that combine FFS with other intervention components, such as input or marketing support, though these studies will be analysed separately.

Types of comparison

If farmers in low and middle income countries do access agricultural extension services at all, it is usually through visits from public extension agents, through observation of public demonstration plots, or through extension provided by the private sector. Public extension may take the form of centralised or more decentralised systems (Birner et al., 2006). We will include studies which compare farmers receiving FFS education to those who receive no, or other types of extension, including where they receive IPM (or equivalent) training from another source. We will collect relevant information on the intervention received by control/comparison groups. Due to the possibilities of spillovers, we will assess whether comparisons are geographically separated from intervention groups, which usually means they are living in a different village to the farmer field school participants.[8]

Types of outcome measures

Primary outcomes

The review primarily looks at economic outcomes, including agricultural yields (production per unit of land), profits (revenues minus costs), household income/ expenditure/ poverty status, and empowerment outcomes such as feelings of self-esteem. The review is interested in effects on two groups of beneficiaries: those participating directly in the field school and those living or working in close proximity to participants (so-called FFS 'exposed'). We will collect information from studies assessing outcomes for either type of beneficiary, and will compare outcomes for both to the non-FFS comparison group.

Secondary outcomes

Intermediate outcomes include farmer knowledge and capacity, adoption of new approaches (including reduced pesticides use) and diffusion of new approaches to 'exposed' farmers who

[8] Ideally, we would include only comparison groups which are separated. However, given problems in operationalising this criteria (most studies reviewed so far do not report sampling procedures), we will instead include all comparisons, assessing likelihood of spillover effects in risk of bias analysis, and reporting effects for separate groups as relevant (e.g. treatment versus control, exposed group versus control).

may live in the same communities as field school graduates, or interact with them at market.

We will also collect data on other final outcomes measured including health and environmental outcomes. These include, for example, self-reporting of health conditions such as respiratory infections or eye irritation, or indices of environmental impact based on assessment of active ingredients in pesticides (Kovach et al., 1992).

Other outcomes/ data

To answer review question (2) we will include data on barriers and facilitators to FFS effectiveness and sustainability. This will include factual information on participation rates and follow up activities and measures of beneficiaries' attitudes and experiences with FFS. It will also include process and implementation information.

Study design and methods of analysis

Review question (1): What is the impact of farmer field schools?

Assessing the measured impacts of FFSs requires an appropriate evaluation methodology. However, designing impact evaluations of agricultural programmes is complicated by the wide range of additional (confounding) factors that influence agricultural outcomes and by biases caused by self-selection of individuals and communities into programmes, meaning that differences in outcomes between participants and non-participants might result from pre-existing differences rather than the programme under evaluation (Romani, 2003).

These problems arise in attempts to attribute the impact of farmer field school programmes on agricultural outcomes. For instance, as Feder et al. (2010) note "the selection of participants into the training is done with strong community involvement through its established leadership and existing social structures" (p. 10). This means that certain farmers, such as community leaders or those of relatively high socio-economic status, may be more likely to benefit from the intervention directly than others, creating difficulties in establishing comparison groups with sufficiently similar characteristics. In addition, pilot programmes may be explicitly placed where they are likely to have the greatest impact. And explicit programme objectives – indeed, ones that may be important for sustainability – are that benefits spillover from participants to non-participants who are 'exposed' to the message through geographical proximity or social networks. In other words, the unit of assessment should be at the community rather than the household level.

In the case of evaluating impacts on agricultural outcomes, such as yields and incomes, the likelihood of serious confounding, particularly by weather and market prices, means that appropriate methods of addressing the attribution challenge necessarily involve comparison groups. However, some might argue that impact evaluations drawing only on pre-intervention and post-intervention data in the intervention group would be appropriate for intermediate outcomes of interest, such as knowledge or even adoption of new technologies,

particularly where it is unlikely beneficiaries would know about them otherwise, as in the case of complex messages. In the case of adoption, farmer behavior is influenced by a range of factors, including policy changes. Removal of subsidies and banning of certain pesticides, as happened in Indonesia in the late 1980s (Braun & Duveskog, 2008) are examples of factors which would likely influence farmers' pesticide behaviour, and in such contexts a before versus after evaluation would not enable researchers to attribute changes to any specific extension interventions. Similarly, farmers might gain knowledge from several places, including public information campaigns, other farmers and other extension interventions. For instance, in Vietnam a 'no early spray' media campaign was run at the same time as an FFS programme, and a study comparing farmers who were exposed to the media campaign with those also attending FFS and a comparison group not exposed to either intervention found that beliefs about insecticide spraying changed in both groups and that the two interventions appeared complementary (Huan et al., 1999). In this case, a simple before versus after (pre-test post-test) comparison would have underestimated the impact of the FFS programme. Finally, while we are interested in capacity building and adoption for their role in the causal chain, our primary interest is in measuring improvements in standards of living.

Therefore, studies eligible for inclusion in the quantitative synthesis must use experimental or quasi-experimental study designs. Study designs which collect longitudinal data at baseline and endline and those using cross-sectional (endline) data only will be included. In addition, data should be collected at the farm or household level contemporaneously in both groups. Studies that use the following allocation methods will be eligible:

- allocation rules based on randomised or quasi-randomised assignment (experimental approaches, RCTs and quasi-RCTs)

- assignment based on other known allocation rules, including a threshold on a continuous variable (regression discontinuity designs, RDDs) or exogenous geographical variation in the treatment allocation ('natural experiments').

- assignment to treatment based on other rules, including self-selection by programme planners and, or participants, provided data is collected in a comparison group (non-equivalent comparison group design), or at least 3 data points are collected both before and after a discrete intervention (six-period interrupted time series).

We will include studies which use statistical matching (e.g. propensity score matching, PSM, or covariate matching), regression adjustment (e.g. difference-in-differences, DID, and single difference regression analysis, instrumental variables, IV, estimation and 'Heckman' selection models), as well as other cross-sectional or longitudinal designs which use less rigorous approaches. Examples of studies included are as follows: Feder et al. (2004a) and Wu (2010) apply multivariate difference-in-differences regression estimation to longitudinal data in Indonesia and China respectively; Ricker-Gilbert et al. (2008) employ an

instrumental variables analysis for cross-sectional data in Vietnam; and van de Fliert (2000) which presents a raw comparison of outcomes across participants and non-participant groups. Given the breadth of designs included, we will conduct rigorous assessment of internal validity and statistical conclusion validity based on risk of bias categories (see below).

Excluded studies are those which do not use a comparison group design, or employ less than a six-period ITS design. For examples, Tin (2009) uses a pre-test post-test design with no comparison group, and Armen et al. (2009) collects post-test data among field school participants only.

Review question (2): Under which circumstances and why?

Studies eligible for inclusion in the synthesis of evidence answering question (2) include any background programme/project documentation, project completion reports and process evaluations which we are able to obtain on the interventions evaluated in the effectiveness studies.

Additionally we will include studies which use quantitative, qualitative or mixed methods of analysis that:

1. report on FFS interventions implemented in the same context (country) as those studies included in the effectiveness synthesis

2. are based on primary data collected from clients, extension agents or experts

3. assess determinants of service delivery quality, knowledge acquisition, adoption of technological improvements, diffusion, or sustainability

4. report at least some information on all of the following: the research question, procedures for collecting data, sampling and recruitment, and at least two sample characteristics.

We will adopt a two-stage approach to inclusion of these studies, which, in addition to removing studies based on the usual relevance criteria (intervention, population, relevance to research question, study type and location), removes studies of particularly low quality in the first round, using the criteria set out in point 4 above. Assessments of quality are then made using a more detailed quality appraisal checklist in the second round, as described below.

Search methods for identification of studies

Electronic searches

We will search a range of different databases, including general social science databases and

subject specific data bases covering agriculture. We will cover the following databases: AgEcon, CAB Abstracts, Social Science Citation Index (SSCI), International Bibliography of Social Science, EconLit, US National Agricultural Library, JOLIS, BLDS, IDEAS and the 3ie impact evaluation database.

To ensure maximal coverage of unpublished literature, we will also search Google and Google Scholar. We will also search the Networked Digital Library of Theses and Dissertations Index to Theses and the ProQuest dissertation database, adapting the search strategy for each database.

Searching the social science literature can be challenging as it is not as well indexed as the medical literature. We developed the search strategy based on the guidance provided in Hammerstrøm et al. (2010), in addition to pearl-harvesting – collecting keywords from studies that meet inclusion criteria (Sandieson, 2006). We will use the following basic search strategy, adapted for each database to include thesaurus terms where available:

'farmer* field* school*' OR ('integrated' AND 'management') AND ('field* school*' or 'farmer* field*')

Other searches

We will screen the bibliographies of included studies and existing reviews for eligible studies. We will hand-search the following journals:

- American Economic Review: applied economics
- Agricultural Economics
- American Journal of Agricultural Economics
- Indian Journal of Extension Education
- Indian Journal of Agricultural Economics
- Integrated Pest Management Reviews
- Journal of Agricultural and Food Information
- Journal of Agricultural Education
- Journal of Agricultural Education & Extension
- Journal of Development Economics
- Journal of Development Studies
- Journal of Development Effectiveness
- Journal of Environmental Extension

- Journal of Extension
- Journal of Extension Systems (India)
- Journal of International Agricultural and Extension Education (AIAEE)
- Journal of International Development
- Journal of Sustainable Agriculture
- Pest management
- SPORE (Netherlands)
- Review of Agricultural Economics
- World Development
- AgBioForum 2009 Special Issue: Herbicide resistant crops
- LEISA 2003 Special issue: Learning with Farmer Field Schools.

We will also conduct forward citation-tracking of included studies in SSCI and Google Scholar.

Finally, we will identify and contact key researchers and organisations working in the field of agricultural extension, including IFAD, IFOAM, ICARDA, Agricultural program IFAP, Environment and Society (Essex University), ODI (Agricultural research and extension network), IEG, IDRC, CGIAR research centres (including IFPRI), FAO, Inmasp, Global IPM Facility, Poverty Action Lab, World Bank, Environment for Development, Practical Action, Oxfam, Farm Africa, and key bilateral donors.

When we have determined which studies will be included in the review of effectiveness we will undertake targeted searching for process and implementation information for those interventions evaluated in the included studies.

Titles and abstracts will be screened against the inclusion criteria and relevant records will be downloaded into the reference management software Refworks. The initial records search will be conducted by two reviewers, screening the records from different databases, and will be over-inclusive to ensure relevant studies are not omitted because sufficient information is not reported in title or abstract. Two reviewers will then independently review downloaded abstracts in more detail to determine which papers should be retrieved and reviewed at full text. Two reviewers will then independently assess full text studies for inclusion, with any disagreements determined by a third reviewer. We will record the details of all searches and report this in the review report.

Data collection and analysis

Selection of studies

Two independent reviewers will assess the full text papers against the inclusion criteria, and each author will extract data from included studies. Discrepancies will be resolved by consensus or by a third author if needed.

Data extraction and management

Two reviewers will extract data from included studies. For the effectiveness synthesis, data will be entered into Excel. We will use nVivo for the barriers and facilitators synthesis.

Studies addressing review question (1)

The following data, where available, will be extracted for each included study: bibliographic information, intervention design, study design, study quality, methods of data collection, context, population, outcomes, explicit or implicit programme theory, information on process and implementation. Fuller details of coding categories are presented in Appendix 1.

Data on effects will be collected from each reference selected for review including on computation procedure of the outcome variable, and the estimated effect and 95 per cent confidence interval. Information will be collected on additional agricultural interventions (for example, access to inputs such as credit) carried out simultaneously.

Studies addressing review question (2)

The following data, where available, will be extracted for each included study: Bibliographic information, intervention design, study design, study quality, methods of data collection, context, population, outcomes, explicit or implicit programme theory, information on process and implementation.

A matrix will be constructed identifying determinants or facilitators of, and barriers to, intervention effectiveness, organised under the preliminary themes of service delivery quality, knowledge acquisition, farmer-to-farmer diffusion, adoption of technology and sustainability. Other themes emerging from the primary studies will also be included.

Quality appraisal

Review question (1): Assessment of risk of bias in included studies of effects

Studies will be critically appraised according to the likely risk of bias based on: 1) quality of attribution methods (addressing confounding and sample selection bias); 2) the extent of

spillovers to farmers in comparison groups;[9] 3) outcome and analysis reporting bias; and 4) other sources of bias. 'Low risk of bias' studies are those in which clear measurement of and control for confounding was made, including selection bias, where intervention and comparison groups were described adequately (in respect of the nature of the interventions being received) and risks of spillovers or contamination were small, and where reporting biases and other sources of bias were unlikely. Studies will be identified as at 'medium risk of bias' where there were threats to validity of the attribution methodology, or there were likely risks of spillovers or contamination, arising from inadequate description of intervention or comparison groups or possibilities for interaction between groups such as when they are from the same community, or reporting biases suspected. 'High risk of bias studies' are all others, including those where comparison groups are not matched or differences in covariates are not accounted for in multivariate analysis, or where there is evidence for spillovers or contamination to comparison groups from the same communities, and reporting biases are evident. Our evaluation criteria are presented in Appendix 2.

Review question 2: Quality appraisal of studies examining barriers and facilitators

We will assess the quality of included studies using an adapted version of the Critical Appraisal Skills Programme checklist (CASP, 2006), making judgments on the adequacy of reporting, data collection, presentation, analysis and conclusions drawn. The checklist is included in Appendix 3. We will filter out studies of particularly low quality at this stage (Noyes et al., 2011) and studies where questions 1-5 are assessed as "No" will be excluded at this stage, and we will not continue with the quality appraisal of such studies.

The remaining studies will be classified as of high or low quality. The results of the quality appraisal will be reported in the review and we will conduct a sensitivity analysis to assess how sensitive our findings are to the removal of studies of different quality (Noyes et al., 2011).

Measures of treatment effect

We will extract comparable effect size estimates from included studies, together with 95 percent confidence intervals. Where possible, we will calculate standardised mean differences (SMDs) for continuous outcome variables, and risk ratios (RRs) for dichotomous outcome variables. Treatment effects will be calculated as the ratio of, or difference between, treated and control observations in a consistent way, such that outcome measures are comparable across studies. Thus, a SMD greater than zero (RR greater than 1) will indicate an increase in the outcome under the intervention as compared to the comparison. A SMD less than zero (RR between 0 and 1) will indicate a reduction under the intervention as

[9] Note that, in contrast, spillovers to 'exposed' farmers are desirable for the intervention, and will be assessed by the measured effects reported on these groups, in separate meta-analysis.

compared to the comparison. A SMD equal to (or insignificantly different from) zero (RR equal to 1) will indicate no change in outcome over the comparison. Whether these relative changes represent positive or negative impacts will depend on meaning of the outcome in the context of the programme being evaluated. For example, while positive impacts on agricultural yields will be measured as values greater than 1, positive impacts of FFSs on – in this case, reductions in – pesticide use will be measured as values less than 1.

We will only include one effect estimate per study in a single meta-analysis. Where studies report multiple effect sizes according to sub-groups of participants, we will report data on sub-groups separately (as in the case of FFS-participant and 'exposed' groups).

Statistical transformations for calculating risk ratios and standardised mean differences from matching-based and regression studies are provided in IDCG (2012).

Unit of analysis issues

The unit of analysis error arises when the unit at which the intervention is implemented and the unit of analysis is different one from each other, for example, when the intervention is delivered at a cluster level (e.g. village or household) but the analysis of impact is carried out at a different unit level (e.g. individual). If clustering in the treatment allocation is not taken into account, the analysis at the individual level would yield narrower confidence intervals increasing the risk of type I error. In other words, if the confidence intervals predicted are narrower than the true confidence intervals, the risk for obtaining a significant result when the true impact is indeed non-significant is larger.

The idea behind the unit of analysis error lies on the assumption that individuals within the same clusters are likely to be more similar than individuals across clusters. In such a case, the observations within clusters cannot be considered independent from one another and therefore the effective sample size is smaller than the total sample size. If the clustering in the treatment assignment is not taken into account in the analysis and the total sample size rather than the effective sample size is used, the confidence intervals would be narrower than the true confidence interval.

In cluster randomised trials with analysis at the individual level, where the assignment to the intervention is carried out across geographical clusters and the variation in the response to the treatment among clusters exceeds the variation within clusters (e.g. individuals or circumstances affecting outcomes within clusters are more similar than across clusters), it is not possible to use the total number of individuals to estimate confidence intervals for an impact effect. This would yield false narrow intervals. Higgins and Green (2011) suggest different solutions for these cases, including analysis at a cluster level or the use of generalized estimating equations. Although the unit of analysis problem has been mainly analysed in the context of cluster randomised trials, it can be also a matter of concern in quasi-experimental studies in which treatment allocation is clustered. Indeed, a systematic

review by Grimshaw et al. (2004) includes an assessment for unit of analysis error in non-randomised controlled before and after studies.

For clustered quasi-experimental studies that use a regression framework, the unit of analysis error arises when, conditional on the covariates and characteristics controlled for, the observations within clusters cannot be considered independent one from each other. That is, when the covariates and methods used in the regression do not fully account for the differences between individuals across clusters. In such a case, the effective sample size is smaller than the total sample size and if this is not taken into account in the design, predicted confidence intervals would be also narrower than the true intervals increasing the risk of falsely rejecting the null of no effect.[10]

This study will provide an assessment of the unit of analysis error for the included studies based on whether the included studies account for differences in demographic and socio-economic household and village characteristics across individuals in different clusters. For those studies that report moderate probability of relevant unit of analysis error, corrections will be applied to the standard errors and the confidence intervals of the effect size using ICCs reported by primary studies if available, or where not available, an intra-cluster correlation coefficient (ICC) of 0.02.[11]

Dealing with missing data

Most quasi-experimental studies used in impact evaluation in economics do not report the information required to calculate standardized mean differences. Where sufficient data are not available to calculate effect sizes, we will contact primary authors to obtain this. Where primary authors are unable to provide relevant information, we will calculate response ratios which offer greater possibilities for estimation. Response ratios measure the proportional change in an outcome in the situation in the intervention group relative to that in the

[10] The validity of regression analysis is based on the assumption of independence of the error term across observations conditional on the covariates. If this condition is not fulfilled, the regression framework yields a biased result. Therefore, in a regression analysis, the existence of unit of analysis error not taken into account in the analysis would not only cause the size of the confidence intervals to be underestimated but also lead to a biased estimate of the impact of the programme.

[11] Several studies (e.g. Campbell et al., 2000; Ukoumunne et al., 1999) report external estimates of ICC for several outcomes. The Health Service Research Unit website includes a spreadsheet with external ICC for different outcomes. Unfortunately, to the best of our knowledge, no previous studies have reported ICC for similar interventions and outcomes. Campbell et al. 2000 suggest that overall, the ICC for outcome variables is not bigger than 0.05 in cluster randomised controlled trials and 0.02 is the mean.

comparison group, giving a similar interpretation to a risk ratio.[12] Borenstein et al. (2009) define this as:

$$R = X_t / X_c$$

where R is the response ratio effect size, X_t is the mean outcome in the treatment group and X_c is the mean outcome in the comparison group. The response ratio provides a measure of the relative change in an outcome caused by an intervention. In other words, the response ratio quantifies the proportionate change that results from an intervention. This systematic review may include different study designs that assess the impact on different measures of the same outcome. For example, studies using a difference-in-differences approach would provide the impact of the programme on the growth rate of the outcome. Other studies that use a propensity score matching approach would provide the impact of the programme on the level of the outcome. Since the response ratio measures the proportional change in an outcome of an intervention, it does not seem unreasonable to combine the response ratios of studies measuring impacts of an intervention on levels with studies assessing impacts on growth rates of outcomes.[13]

Assessment of heterogeneity

We will report tests for heterogeneity of effects across studies, using the I^2 statistic and by reporting the between studies variance component (τ^2) to provide an overall estimate of the amount of variability in the distribution of the true effect sizes (Borenstein et al., 2009).

Assessment of reporting biases

For dependent effect sizes, where multiple outcome measures are reported by sub-group or when the impact of the programme on multiple outcomes measuring the same outcome category are reported, we will combine groups from the same study prior to meta-analysis, in

[12] The use of response ratios to combine study results is subject to some limitations. Borenstein et al. (2009) highlight that this effect size is only meaningful when the outcome is measured on a true ratio scale that has a natural zero point (though is unlikely to be equal to zero in practice). This condition holds for the outcomes measured here (knowledge scores as percentage of 'correct' answers, pesticide use and costs, production yield and revenues, disease incidence).

[13] On the other hand it would not be meaningful to combine standardized mean differences or mean differences of studies measuring impact in yields levels with studies measuring impact on growth rate of yields. Indeed, the mean differences approaches might require included studies to use not only the same outcome but also the same measure of outcome, preventing the aggregation of results of studies that use study designs based on panel data (cross-sectional before versus after) and those based on cross-sectional data only.

order to avoid accusations of results-related choices. Where studies report multiple effects due to sub-groups of participants or, for example, follow up period, we will synthesise effects prior to meta-analysis. We will calculate the sample-weighted average effect sizes, using appropriate formulae to recalculate variances, making necessary covariance assumptions if necessary (Borenstein et al., 2009; Higgins & Green, 2011, Chapter 16). Where multiple outcomes are reported from alternate specifications, we will select the specification according to likely lowest risk of bias in attributing impact.

Methods of synthesis

The review synthesises quantitative data on effectiveness to assess whether the intervention works or not (objectives question 1), and mixed (quantitative and/or qualitative) data on barriers and facilitators to explain why (objectives question 2).

Review question (1): Effectiveness synthesis

Effect sizes will be reported using forest plots and synthesised in inverse-variance weighted meta-analysis, estimated using Stata software (Stata Corporation, College Station, TX, USA). Owing to contextual heterogeneity, random effect meta-analysis will be employed. By accounting for the possibility of different effect sizes across studies in this way, random effects meta-analysis produces a pooled effect size with greater uncertainty attached to it, in terms of wider confidence intervals than a fixed effect model.

Review question (2): Barriers and facilitators synthesis

For the synthesis of evidence relating to question (2), we will use a thematic approach, where themes will be based on the links and assumptions in the theory of change model. The choice method of synthesis of data relating to barriers and facilitators depends on the type and scope of the review question, the type of evidence in included studies and the degree to which there are existing theoretical frameworks for the issue under review in the literature (Noyes et al., 2011). The scope of this component of the review is closely related to the effectiveness question and a preliminary review of the evidence suggests that the included studies are likely to be descriptive, or 'thin' qualitative studies. An integrative/ aggregative synthesis approach is therefore likely to be most appropriate, and we anticipate using a combination of content analysis and narrative synthesis.

Integrated synthesis (review questions 1 and 2)

We will use the program theory as a framework for integrating the findings from synthesis of review questions (1) and (2) in a narrative synthesis along the causal chain. The question addressed in synthesis (2) is directly related to, and aims to enhance the findings of the effectiveness review. Hence the synthesis will integrate the findings from the two syntheses (Noyes et al., 2011) with the aim of providing an integrated narrative synthesis addressing the objectives of the review. The implications of these findings for programme planning will

then be discussed.

Subgroup analysis and investigation of heterogeneity

We will investigate whether findings differ according to key contextual factors, including geographical region, crop type, length of exposure (measured as length of FFS programme implementation, and length of post-implementation follow-up period). We will also collect information on sub-groups of interest according to PROGRESS-plus categories, including women and other vulnerable groups.

Sensitivity analysis

We will conduct sensitivity analysis according to categories of risk of bias, study design (experimental and quasi-experimental), treatment effect (e.g. intention to treat, average treatment effect on the treated, local average treatment effect), and also in the event of outliers.

While we attempt to reduce publication bias by searching for and including unpublished studies in the review, we will also test for possible publication bias using meta-analysis, and for under-reporting of small sample studies using funnel plots and Egger et al.'s (1998) test. Given the inherent subjectivity in assessing funnel plot asymmetry, we will assess sensitivity of meta-analyses using 'trim and fill' (Duvall & Tweedie, 2005) regardless of whether funnel plots suggest asymmetry.[14]

[14] The trim-and-fill method does not allow for moderators to be included in the model; where data allow, we will therefore consider other methods of publication bias analysis, including Vevea and Hedges' (1995) weight-function model.

REFERENCES

3ie (n.d.). *Principles for impact evaluation.* International Initiative for Impact Evaluation: New Delhi. Available at http://www.3ieimpact.org/strategy/pdfs/principles for impact evaluation.pdf

Akroyd, S. & Smith, L. (2007). The decline in public spending to agriculture – does it matter? *Briefing Note, No. 2*. Oxford Policy Management Institute: Oxford.

Alston, J. M., Wyatt, T. J., Pardey, P. G., Marra, M. C. & Chan-Kang, C. (2000). *A meta-analysis of rates of return to agricultural R&D – Ex Pede Herculem*. IFPRI: Washington D.C.

Anderson, J. R. (2007). Agricultural advisory services. *Background Paper for the World Development Report 2008*. Available at http://siteresources.worldbank.org/INTWDR2008/Resources/2795087-1191427986785/Anderson_AdvisoryServices.pdf

Armen, V., Hanson, J., & Houseman, I. (2009) Assessing the impact of the agricultural advisory systems in Armenia. In C. Paffarini and F.M. Santucci (Eds.) *19th European seminar on extension education: Theory and practice of advisory work in a time of turbulence*. Perugia: DSEEA, Facoltà di agraria.

Becker, B., Hedges, L., & Pigott, T. (n.d.). *Statistical analysis policy brief*. Campbell Collaboration: Oslo. Available at http://www.campbellcollaboration.org/artman2/uploads/1/C2_Statistical_Analysis_Policy_Brief-2.pdf

Birkhaeuser, D., Evenson, R. E., & Feder, G., (1991). The economic impact of agricultural extension: A review. *Economic Development and Cultural Change, 39*, 607-650.

Birner, R., Davis, K., Pender, J., Nkonya, E., Anandajayasekeram, P., Ekboir, J., ... Cohen, M. (2006). From "best practice" to "best fit": A framework for analyzing pluralistic agricultural advisory services worldwide. *DSGD Discussion Paper No. 37*. IFPRI: Washington D.C. Available at http://www.ifpri.org/DIVS/DSGD/dp/dsgdp37.asp.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons, Ltd: Chichester, UK. doi: 10.1002/9780470743386.refs

Braun, A. R., Thiele, G., & Fernandez, M. (2000). Farmer field schools and local agricultural research committees: complementary platforms for integrated decision-making in sustainable agriculture. *Agricultural Research and Extension Network Paper No.105*. Overseas Development Institute.

Braun, A., & Duveskog, D. (2008). The Farmer Field School approach – History, global assessment and success stories. *Background Paper for the IFAD Rural Poverty Report 2010*.

Braun, A., Jiggins, J., Röling, N., van den Berg, H., & Snijders, P. (2006). A global survey and review of Farmer Field School experiences. *Report prepared for the International Livestock Research Institute (ILRI)*. Available at http://intranet.catie.ac.cr/intranet/posgrado/Met%20Cual%20Inv%20accion/MCIAP2010/Semana%203/DocumentosSem310/Review%20of%20FFS%20Braun%202006.pdf.

Campbell, M., Grimshaw, J., & Steen, N. (2000). Sample size calculations for cluster randomised trials. *Journal of Health Services Research and Policy, 5*, 12-16.

Critical Appraisal Skills Programme (CASP). (2006). *10 questions to help you make sense of qualitative research.* Public Health Resource Unit: England. http://www.phru.nhs.uk/Doc_Links/Qualitative%20Appraisal%20Tool.pdf

Davis, K. (2006). Farmer Field Schools: A boon or bust for extension in Africa? *Journal of International Agricultural and Extension Education, 13*(1), 91-97.

Davis, K., Nkonya, E., Kato, E., Mekonnen, D. A., Odendo, M., Miiro, R., ... Okoth, J. (2010). *Impact of farmer field schools on agricultural productivity and poverty in East Africa*. IFPRI: Washington D.C.

Diabate, A., Baldet, T., Chandre, F., Akogbeto, M., Guiduemde, T.R., Darriet, F., ... Hougard, J. M. (2002) The role of agricultural use of insecticides in resistance to pyrethroids in Anopheles Gambiae s.l. in Burkina Faso. *American Journal of Tropical Medicine and Hygiene, 67*, 617-622.

Duval, S., & Tweedie, R. (2000) A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89-98.

Effective Practice and Organisation of Care Group (EPOC). (n.d.). *Suggested risk of bias criteria for EPOC reviews.* Available from http://epocoslo.cochrane.org/epoc-specific-resources-review-authors

Egger, M., Davey Smith, G., Schneider, M. & Minder, C. (1997) Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634.

Fan, S., & Rao, N. (2003). Public spending in developing countries: Trends, determination, and impact. *EPTD Discussion Paper, No. 99*. IFPRI: Washington D.C. Available at http://ideas.repec.org/p/fpr/eptddp/99.html

FAO. (2003). Statement at the ministerial conference of the WTO. *Circulated by H.E. Mr Hartwig de Haen, Assistant Director-General', Fifth Session, Cancun, 10-14 September*. Doc No: WT/MIN(03)/ST/61.

Feder, G., Murgai, R., & Quizon, J. B. (2004a). Sending farmers back to school: The impact of farmer field schools in Indonesia. *Review of Agricultural Economics, 26*(1), 45-62.

Feder, G., Murgai, R., & Quizon, J. B. (2004b). The acquisition and diffusion of knowledge: The case of pest management training in farmer field schools, Indonesia. *Journal of Agricultural Economics, 55*(2), 221-243.

Feder, G., & Savastano, S. (2006). The role of opinion leaders in the diffusion of new knowledge: The case of integrated pest management. *World Bank Policy Research Working Paper 3916*. World Bank: Washington D.C. Available at http://econpapers.repec.org/paper/wbkwbrwps/3916.htm

Feder, G., Anderson, J. R., Birner, R., & Deininger, K. (2010). Promises and realities of community-based agricultural extension. *IFPRI Discussion Paper 00959*. Washington D.C.: International Food Policy Research Institute.

Gautam, M., & Anderson, J. (2000). Agricultural extension: The Kenya experience: An impact evaluation. *Operations Evaluation Department*. World Bank: Washington, D.C.

Green, D., Lee, B., Morrison, J., & Werth, A. (2005). Sustainable development, poverty and agricultural trade reform. Chapter 2 in *Agricultural commodities, trade and sustainable development*. International Institute for Environment and Development: London.

Hammerstrøm, K., Wade, A., & Klint Jørgensen, A.-M. (2010) *Searching for studies: A guide to information retrieval for Campbell Systematic Reviews*. Campbell Collaboration: Oslo. Available at www.campbellcollaboration.org/lib/download/969/

Higgins, J., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions. (Version 5.0.2, updated September 2009).* The Cochrane Collaboration. Available at www.cochrane-handbook.org

Huan, N. H., Mai, V., Escalada, M. M., & Heong, K. L. (1999). Changes in rice farmers' pest management in the Mekong Delta, Vietnam. *Crop Protection, 18*, 557-563.

International Development Coordinating Group (IDCG). (2012). *Protocol and review guidelines*. Campbell Collaboration. Available at www.campbellcollaboration.org

International Fund for Agricultural Development (IFAD). (2001). *Rural Poverty Report*

2011: The Challenge of ending rural poverty. Rome. Available at http://www.ifad.org/poverty/

International Fund for Agricultural Development (IFAD) (2010) Rural Poverty Report 2011: New realities, new challenges: new opportunities for tomorrow's generation, Rome. http://www.ifad.org/rpr2011/report/e/rpr2011.pdf

Inter-Academy Council. (2004). *Realizing the promise and potential of African agriculture: Science and technology strategies for improving agricultural productivity and food security in Africa*. InterAcademy Council, Amsterdam.

Kelly, L. (2005) *The global integrated pest management facility: Addressing challenges of globalization: An independent evaluation of the World Bank's approach to global programs, case study*. Operations Evaluation Department. World Bank: Washington, DC. Available at http://lnweb90.worldbank.org/oed/oeddoclib.nsf/DocUNIDViewForJavaSearch/210300C07054C81A852570A50076C0DC/$file/gppp_pest_management_facility.pdf

Khisa, G. (2004). *Farmers Field School methodology: Training of trainers manual (1st Edition)*. FAO, Rome.

Kovach, J., Petzoldt, C., Degni, J., & Tette, J. (1992). A method to measure the environmental impact of pesticides. *New York's Food and Life Sciences Bulletin, 139*, 1-8.

Munro, S. A., Lewin, S. A., Smith, H. J., Engel, M. E., Fretheim, A., & Volmink, J. (2007) Patient adherence to tuberculosis treatment: A systematic review of qualitative research. *PLoS Med, 4*(7): e238. doi:10.1371/journal.pmed.0040238.

Noyes, J., Booth, A., Hannes, K., Harden, A., Harris, J., Lewin, S., & Lockwood, C. (Eds.). (2011). *Supplementary guidance for inclusion of qualitative research in Cochrane systematic reviews of interventions (Version 1, updated August 2011)*. Cochrane Collaboration Qualitative Methods Group. Available from http://cqrmg.cochrane.org/supplemental-handbook-guidance

Pingali, P. L., & Roger, P. A. (1995). *Impact of pesticides on farmer health and the rice eco-system*. Kluwer Academic Publishers: Boston.

Pontius, J., Dilts, R., & Bartlett, A. (2002). *From Farmer Field School to Community IPM: ten years of IPM training in Asia*. Food and Agriculture Organiszation of the United Nations, FAO Regional Office for Asia and the Pacific: Rome.

Quizon, J., Feder, G., & Murgai, R. (2001) Fiscal sustainability of agricultural extension: The case of the Farmer Field School approach - supplementary remarks. *Journal of*

International Agricultural and Extension Education. Available at http://www.aiaee.org/attachments/303_Quizon-Vol-8.3-7.pdf

Ricker-Gilbert, J., Norton, G. W., Alwang, J., Miah, M. & Feder, G. (2008). Cost-effectiveness of alternative integrated pest management extension methods: An example from Bangladesh. *Review of Agricultural Economics, 30*(2), 252-269.

Romani, M. (2003). *The impact of extension services in times of crisis: Cote d'Ivoire (1997-2000)*. CSAE WPS/2003-07, Centre for the Study of African Economies: University of Oxford. Available at http://ideas.repec.org/p/wpa/wuwpdc/0409053.html

Sandieson, R. (2006) Pathfinding in the research forest: The pearl harvesting method for effective information retrieval. *Education and Training in Developmental Disabilities 41*(4), 401-409. Available at http://publish.edu.uwo.ca/robert.sandieson/downloads/ETDD.pdf

Shadish, W. and Myers, D. (2004). Research design policy brief. Campbell Collaboration: Oslo. Available at http://www.campbellcollaboration.org/artman2/uploads/1/C2_Research_Design_Policy_Brief-2.pdf

Shemilt, I., Mugford, M., Byford, S., Drummond, M., Eisenstein, E., Knap, M., ... Walker, D. (2008). *The Campbell Collaboration Economics Methods Policy Brief*. Campbell Collaboration: Oslo. Available at http://www.campbellcollaboration.org/artman2/uploads/1/Economic_Methods_Policy_Brief.pdf

Stathers, T., Namanda, S., Mwanga, R. O. M., Khisa, G., & Kapinga, R. (2005) *Manual for Sweetpotato Integrated Production and Pest Management Farmer Field Schools in Sub-Saharan Africa*. International Potato Center: Kampala, Uganda.

Tin, H. Q. (2009) *Impacts of farmer-based training in seed production in Vietnam*. Thesis. Wageningen University.

Tripp, R., Wijeratne, M., & Piyadasa, V. H. (2005). What should we expect from Farmer Field Schools? A Sri Lanka case study. *World Development, 33*(10), 1705-1720.

White, H. (2009) Theory-based impact evaluation: Principles and practice. *Working Paper 3*. International Initiative for Impact Evaluation: New Delhi. Available at http://www.3ieimpact.org/admin/pdfs_papers/51.pdf

Ukoumunne, O. C., Gulliford, M. C., Chinn, S., Sterne, J. A., & Burney, P. G. Methods for evaluating area-wide and organisation-based interventions in health and health care: A systematic review. *Health Technology Assessment, 3*: 5.

United Nations. (2008). *Trends in sustainable development – Agriculture, rural development, land, desertification and drought*. Department of Economic and Social Affairs of the United Nations: New York. Available at http://www.un.org/esa/sustdev/publications/trends2008/fullreport.pdf

Van de Fliert. (2000). Stepping stones and stumbling blocks in capacity development of Sweetpotato ICM Farmer Field School facilitators. *UPWARD Conference "Capacity Development for Participatory Research", Beijing, 19-22 September*.

Van den Berg, H. (2004). IPM Farmer Field Schools: A synthesis of 25 impact evaluations. *Prepared for the Global IPM Facility, Wageningen University, the Netherlands*. Available at ftp://ftp.fao.org/docrep/fao/006/ad487E/ad487E00.pdf

Van den Berg, H., & Jiggins, J. (2007). Investing in farmers – The impacts of Farmer Field Schools in relation to integrated pest management. *World Development, 35*(4), 663-686.

Vevea, J. L., & Hedges, L.V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*, 419-435.

Waddington, H., Snilstveit, B., White, H., & Anderson, J. (2010). The impact of agricultural extension services: Study protocol. *3ie Synthetic Reviews SR009*. International Initiative for Impact Evaluation: New Delhi. Available at http://www.3ieimpact.org/admin/pdfs_synthetic/009%20Protocol.pdf

Williamson, L. M., Parkes, A., Wight, D., Petticrew, M., & Hart, G. (2009). Limits to modern contraceptive use among young women in developing countries: A systematic review of qualitative research, *Reproductive Health*, *6*, 3 doi:10.1186/1742-4755-6-3.

World Bank. (2007). *World Development Report 2008*. Agriculture for Development, World Bank: Washington, D.C. Available at http://siteresources.worldbank.org/INTWDR2008/Resources/WDR_00_book.pdf

Wu, L. (2010). Farmer Field School and Bt Cotton in China – An economic analysis. In Waibel, H. (Ed.), *Pesticide policy project publication series, special issue No. 15, March 2010*. Institute of Development and Agricultural Economics, Leibniz University of Hannover, Germany.

SOURCES OF SUPPORT

We would like to thank the Millennium Challenge Corporation (MCC) of the US Government and 3ie for financial support, and Sandra-Jo Wilson of the Campbell Education Group for coordinating the peer review.

DECLARATIONS OF INTEREST

We are not aware of any conflicts of interest arising from either researcher interest or financial sources.

REVIEW TEAM

Lead reviewers:

The lead authors are the people who develop and co-ordinate the review team, discuss and assign roles for individual members of the review team, liaise with the editorial base and take responsibility for the on-going updates of the review.

Name:	**Hugh Waddington**
Title:	Senior Evaluation Officer
Affiliation:	International Initiative for Impact Evaluation
Address:	London International Development Centre
City, State, Province or County:	36 Gordon Square, London
Postal Code:	WC1H 0PD
Country:	UK
Phone:	+44 207 958 8352
Mobile:	
Email:	hwaddington@3ieimpact.org

Name:	**Birte Snilstveit**
Title:	Evaluation Officer
Affiliation:	International Initiative for Impact Evaluation
Address:	London International Development Centre
City, State, Province or County:	36 Gordon Square, London
Postal Code:	WC1H 0PD
Country:	UK
Phone:	+44 207 958 8352

Mobile:	
Email:	bsnilstveit@3ieimpact.org

Co-authors:

Name:	**Jorge Garcia Hombrados**
Title:	Research Assistant
Affiliation:	International Initiative for Impact Evaluation
Address:	London International Development Centre
City, State, Province or County:	36 Gordon Square, London
Postal Code:	WC1H 0PD
Country:	UK
Phone:	+44 207 958 8352
Mobile:	
Email:	jhombrados@3ieimpact.org

Name:	**Martina Vojtkova**
Title:	Research Assistant
Affiliation:	International Initiative for Impact Evaluation
Address:	London International Development Centre
City, State, Province or County:	36 Gordon Square, London
Postal Code:	WC1H 0PD
Country:	UK
Phone:	+44 207 958 8352
Mobile:	
Email:	mvojtkova@3ieimpact.org

Name:	**Jock Anderson**
Title:	Consultant
Affiliation:	World Bank
Address:	
City, State, Province or County:	
Postal Code:	
Country:	
Phone:	
Mobile:	

Email:	janderson@worldbank.org

Name:	**Howard White**
Title:	Executive Director
Affiliation:	International Initiative for Impact Evaluation
Address:	London International Development Centre
City, State, Province or County:	36 Gordon Square, London
Postal Code:	WC1H 0PD
Country:	UK
Phone:	
Mobile:	
Email:	hwhite@3ieimpact.org

REQUEST SUPPORT

No support is requested at this time.

ROLES AND RESPONSIBLIITIES

The protocol was developed by Hugh Waddington (HJW) and Birte Snilstveit (BS) with contributions from Jorge Hombrados (JH). BS, HJW, JH, and Martina Vojtkova (MV) will conduct the search, using reference management software. Decisions on inclusion for impact evaluation studies will be made by HJW, BS, and Jorge Hombrados (JH), with conflicts resolved through discussion and consensus. Qualitative study coding will be carried out by BS, MV, and JH, while critical appraisal and effect sizes estimation will be done by JH and HJW. Howard White and Jock Anderson will provide technical support.

PRELIMINARY TIMEFRAME

We expect to submit the review in October 2012.

PLANS FOR UPDATING THE REVIEW

The experimental and quasi-experimental research in this area is very limited and the rate of publication of new high quality studies is likely to be slow. We will keep abreast of the literature in the field and update the review once sufficient high quality studies become available.

AUTHORS' RESPONSIBILITIES

By completing this form, you accept responsibility for preparing, maintaining and updating the review in accordance with Campbell Collaboration policy. The Campbell Collaboration will provide as much support as possible to assist with the preparation of the review.

A draft review must be submitted to the relevant Coordinating Group within two years of protocol publication. If drafts are not submitted before the agreed deadlines, or if we are unable to contact you for an extended period, the relevant Coordinating Group has the right to de-register the title or transfer the title to alternative authors. The Coordinating Group also has the right to de-register or transfer the title if it does not meet the standards of the Coordinating Group and/or the Campbell Collaboration.

You accept responsibility for maintaining the review in light of new evidence, comments and criticisms, and other developments, and updating the review at least once every three years, or, if requested, transferring responsibility for maintaining the review to others as agreed with the Coordinating Group.

PUBLICATION IN THE CAMPBELL LIBRARY

The support of the Campbell Collaboration and the relevant Coordinating Group in preparing your review is conditional upon your agreement to publish the protocol, finished review and subsequent updates in the Campbell Library. Concurrent publication in other journals is encouraged. However, a Campbell systematic review should be published either before, or at the same time as, its publication in other journals. Authors should not publish Campbell reviews in journals before they are ready for publication in the Campbell Library. Authors should remember to include the statement: "This is a version of a Campbell review, which is available in The Campbell Library" when publishing in journals or other venues.

I understand the commitment required to undertake a Campbell review, and agree to publish in the Campbell Library. Signed on behalf of the authors:

Form completed by: Hugh Waddington

Date: 9 September 2012

APPENDIX 1 DATA COLLECTION VARIALBES

General information	Author surname
	Year of publication
	Publication type: journal article, working paper, book
	Funder of intervention
	Author affiliation: employee of implementing agency, employee of other body.
Intervention design	Intervention type: IPM, IPPM, other
	Components of intervention: size of field school group, use of control plot.
	Intervention period (from MM/YY to MM/YY)
	Additional interventions provided: e.g. input support, market support.
	Target group: women, men, both.
Context	Country and region (EAP, LAC, MENA, SA, SSA)
	Crops: cotton, rice, vegetables, other.
	Length of follow-up: months from intervention to outcomes data collection
Study design	Study type: RCT, quasi-RCT, RDD, natural experiment, DID, IV, ITS, PSM, adjusted (multivariate) single difference regression, unadjusted comparison of means
	Description of comparison group (and if relevant 'exposed' group) intervention
	Period of outcomes data collection (from MM/YY to MM/YY)
	Frequency of outcomes data collection
	Information reported on method of allocating individuals to groups
	Sample size (treatment, exposed, comparison): number of clusters, number of individuals
	Sample attrition (treatment, exposed, comparison)
	Spillovers: geographical separation of treatment and comparison
	Contamination: influence of other intervention which differentially affects treatment and comparison groups on relevant outcomes.
Study quality	Risk of bias assessment (see Appendix 2).
Effect estimation	Treatment effect estimated: ITT, ATET, ATE, LATE
	Adjusted or unadjusted analysis.
Intermediate outcomes	Knowledge: e.g. knowledge of 'simple', 'intermediate' and 'complex' practices
	Adoption: e.g. use of fertiliser, cost of fertiliser, number of 'improved' practices.
Final outcomes	Yields: weight per unit of land
	Revenues: value of production minus costs of production (farm income or profits)
	Environment: e.g. Environmental Impact Quotient score

	Health: e.g. incidence of health complaint (eye irritation, respiratory problems, stomach ache) Empowerment: e.g. reported self-esteem.
Qualitative/quantitative information	Barriers to and enablers of: service delivery quality (e.g. approach to training of trainers, field school period timed with growing period), diffusion of information (e.g. method of communication, farmers attend sufficient schooldays), adoption of technological improvements (e.g. farmer time, costs of inputs), sustainability (e.g. community orientation, training of farmer-trainers).

APPENDIX 2 CRITICAL APPRAISAL OF IMPACT STUDIES TO ANSWER RESEARCH QUESTION (1)[15]

1) Selection bias and confounding

a) For Randomised assignment (RCTs),

Score "YES" if:

- a random component in the sequence generation process is described (e.g. referring to a random number table)[16];
- and if the unit of allocation was at group level (geographical/ social/ institutional unit) and allocation was performed on all units at the start of the study,
- or if the unit of allocation was by beneficiary or group and there was some form of centralised allocation mechanism such as an on-site computer system;
- and if the unit of allocation is based on a sufficiently large sample size to equate groups on average.
- baseline characteristics of the study and control/comparisons are reported and overall[17] similar based on t-test or ANOVA for equality of means across groups,
- or covariate differences are controlled using multivariate analysis;
- and the attrition rates (losses to follow up) are sufficiently low and similar in treatment and control, or the study assesses that loss to follow up units are random draws from the

[15] We drew on 3ie (n.d.) and EPOC (n.d.) in developing this tool.

[16] If a quasi-randomized assignment approach is used (e.g. alphabetical order), you must be sure that the process truly generates groupings equivalent to random assignment, to score "Yes" on this criteria. In order to assess the validity of the quasi-randomization process, the most important aspect is whether the assignment process might generate a correlation between participation status and other factors (e.g. gender, socio-economic status) determining outcomes; you may consider covariate balance in determining this (see question 2).

17 Even in the context of RCTs, when randomisation is successful and carried out over sufficiently large assignment units, it is possible that small differences between groups remain for some covariates. In these cases, study authors should use appropriate multivariate methods to correcting for these differences.

sample (e.g. by examining correlation with determinants of outcomes, in both treatment and comparison groups);

- and problems with cross-overs and drop outs are dealt with using intention-to-treat analysis or in the case of drop outs, by assessing whether the drop outs are random draws from the population;

- and, for cluster-assignment, authors control for external cluster-level factors that might confound the impact of the programme (eg weather, infrastructure, community fixed effects, etc) through multivariate analysis.

Score "UNCLEAR" if:

- the paper does not provide details on the randomisation process, or uses a quasi-randomization process for which it is not clear has generated allocations equivalent to true randomisation.

- insufficient details are provided on covariate differences or methods of adjustment;

- or insufficient details are provided on cluster controls.

Score "NO" if:

- the sample size is not sufficient or any failure in the allocation mechanism or execution of the method could affect the randomisation process[18].

b) For discontinuity assignment (regression discontinuity design)

Score "YES" if:

- allocation is made based on a pre-determined discontinuity on a continuous variable (regression discontinuity design) and blinded to participants or,

- if not blinded, individuals reasonably cannot affect the assignment variable in response to knowledge of the participation decision rule;

- and the sample size immediately at both sides of the cut-off point is sufficiently large to equate groups on average.

- the interval for selection of treatment and control group is reasonably small,

[18] If the research has serious concerns with the validity of the randomisation process or the group equivalence completely fails, we recommend to assess the 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 RCTs questions.

- or authors have weighted the matches on their distance to the cut-off point,
- 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,
- or significant differences 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 (eg weather, infrastructure, community fixed effects, etc) through multivariate analysis.

Score "UNCLEAR" if:

- the assignment variable is either non-blinded or it is unclear whether participants can affect it in response to knowledge of the allocation mechanism.
- there are covariate differences across individuals at both sides of the discontinuity which have not been controlled for using multivariate analysis, or if insufficient details are provided on controls,
- or if insufficient details are provided on cluster controls.

Score "NO" if:

- the sample size is not sufficient or
- there is evidence that participants altered the assignment variable prior to assignment[19].

c) For identification based on an instrumental variable (IV estimation)

Score "YES" if:

- An appropriate instrumental variable is used which is exogenously generated: e.g. due to a 'natural' experiment or random allocation.
- the instrumenting equation is significant at the level of≥F10 (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);

[19] If the research has serious concerns with the validity of the assignment process or the group equivalence completely fails, we recommend to assess the 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.

- the identifying instruments are individually significant ($p \leq 0.01$); for Heckman models, the identifiers are reported and significant ($p \leq 0.05$);

- where at least two instruments are used, the authors report on an over-identifying test ($p \leq 0.05$ 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[20].

- and, for cluster-assignment, authors particularly control for external cluster-level factors that might confound the impact of the programme (eg weather, infrastructure, community fixed effects, etc) through multivariate analysis.

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).

- relevant confounders are controlled but appropriate statistical tests are not reported or exogeneity[21] of the instrument is not convincing,

- or if insufficient details are provided on cluster controls (see category f) below).

Score "NO" otherwise.

d) For assignment based non-randomised programme placement and self-selection (studies using a matching strategy or regression analysis (excluding IV), studies which apply other methods)

Score "YES" if:

- Participants and non-participants are either matched based on all relevant characteristics explaining participation and outcomes, or

- all relevant characteristics are accounted for.[22] [23]

---

[20] 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).

[21] An instrument is exogenous when it only affects the outcome of interest through affecting participation in the programme. Although when more than one instrument is available, statistical tests provide guidance on exogeneity (see background document), the assessment of exogeneity should be in any case done qualitatively. Indeed, complete exogeneity of the instrument is only feasible using randomised assignment in the context of an RCT with imperfect compliance, or an instrument identified in the context of a natural experiment.

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 “NO” if:

- relevant characteristics are omitted from the analysis.

In addition:

d1) For non-randomised trials using panel data (including DID) models,

Score “YES” if:

- the authors use a difference-in-differences (or fixed effects) multivariate estimation method;
- the authors control for a comprehensive set of time-varying characteristics;[24]
- 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 (e.g. by examining correlation with determinants of outcomes, in both treatment and comparison groups);

---

[22] 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.

[23] There are different ways in which 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.

[24] 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.

- and, for cluster-assignment, authors control for external cluster-level factors that might confound the impact of the programme (eg weather, infrastructure, community fixed effects, etc) through multivariate analysis.

Score "UNCLEAR" if:

- insufficient details are provided,
- or if insufficient details are provided on cluster controls.

Score "NO" otherwise, including if the treatment effect is estimated using raw comparison of means in statistically un-matched groups.

d2) For statistical matching studies including propensity scores (PSM) and covariate matching,[25]

Score "YES" if:

- 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 (e.g. demographic and socio-economic factors) to explain both participation and the outcome (so that there can be no evident differences across groups in variables that might explain outcomes) (see fn. 6).
- In addition, for PSM Rosenbaum's test suggests the results are not sensitive to the existence of hidden bias.
- and, with the exception of Kernel matching, the means of the individual covariates are equated for treatment and comparison groups after matching;
- and, for cluster-assignment, authors control for external cluster-level factors that might confound the impact of the programme (eg weather, infrastructure, community fixed effects, etc) through multivariate or any appropriate analysis.

Score "UNCLEAR" if:

- relevant variables are not included in the matching equation, or if matching is based on characteristics collected at endline,

[25] Matching strategies are 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 unobservables differences across groups affecting outcome of interest and removing biases arising from time-invariant unobservable characteristics.

- or if insufficient details are provided on cluster controls.

Score “NO” otherwise.

d3) For regression-based studies using cross sectional data (excluding IV)

Score “YES” if:

- the study controls for relevant confounders that may be correlated with both participation and explain outcomes (e.g. demographic and socio-economic factors at individual and community level) using multivariate methods with appropriate proxies for unobservable covariates (see fn. 6),
- and a Hausman test[26] with an appropriate instrument suggests there is no evidence of endogeneity,
- and none of the covariate controls can be affected by participation;
- 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;
- and, for cluster-assignment, authors control particularly for external cluster-level factors that might confound the impact of the programme (eg weather, infrastructure, community fixed effects, etc) through multivariate analysis.

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 “NO” otherwise.

[26] 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.

d4) For study designs which do not account for differences between groups using statistical methods, score "NO".

2) Spill-overs: was the study adequately protected against performance bias?

Score "YES" if:

- 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 unlikely)[27].

Score "UNCLEAR" if:

- spill-overs are not addressed clearly.

Score "NO" if:

- allocation was at individual or household level and there are likely spill-overs within households and communities which are not controlled for in the analysis;
- or if allocation at cluster level and there are likely spill-overs to comparison clusters.

3) Selective reporting: was the study free from outcome and analysis reporting biases?

Score "YES" if:

- there is no evidence that outcomes were selectively reported (e.g. all relevant outcomes in the methods section are reported in the results section).
- authors use 'common' methods[28] of estimation and the study does not suggest the existence of biased exploratory research methods[29].

---

[27] Contamination, that is differential receipt of other interventions affecting outcome of interest in the control or comparison group, is potentially an important threat to the correct interpretation of study results and should be addressed via PICO and study coding.

[28] 'Common methods' refers to the use of the most credible method of analysis to address attribution given the data available.

Score “NO” if:

- some important outcomes are subsequently omitted from the results or the significance and magnitude of important outcomes was not assessed.
- authors use uncommon or less rigorous estimation methods such as failure to conduct multivariate analysis for outcomes equations where it is has not been established that covariates are balanced.[30]

Score “UNCLEAR” otherwise.

4) Other: was the study free from other sources of bias?

Important additional sources of bias may include: concerns about blinding of outcome assessors or data analysts; concerns about blinding of beneficiaries so that expectations, rather than the intervention mechanisms, are driving results (detection bias or placebo effects)[31]; concerns about courtesy bias from outcomes collected through self-reporting;

---

[29] A comprehensive assessment of the existence of ‘data mining’ is not feasible particularly in quasi-experimental designs where most studies do not have protocols and replication seems the only possible mechanism to examine rigorously the existence of data mining.

[30] For PSM and covariate matching, score “YES” if: where over 10% of participants fail to be matched, sensitivity analysis is used to re-estimate results using different matching methods (Kernel Matching techniques). For matching with replacement, no single observation in the control group is matched with a large number of observations in the treatment group. Where not reported, score “UNCLEAR”. Otherwise, score “NO”.

For IV (including Heckman) models, score “YES” if: 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), the coefficient of the selectivity correction term (Rho) is significantly different from zero ($P<0.05$) (Heckman approach). Where not reported, score “UNCLEAR”. Otherwise, score “NO”.

For studies using multivariate regression analysis, score “YES” if: authors conduct appropriate specification tests (e.g. reporting results of multicollinearity test, testing robustness of results to the inclusion of additional variables, etc). Where not reported or not convincing, score “UNCLEAR”. Otherwise, Score “NO”.

[31] All interventions may create expectations (placebo effects), which might confound causal mechanisms. In social interventions, which usually require behaviour change from

concerns about coherence of results; data on the baseline collected retrospectively; information is collected using an inappropriate instrument (or a different instrument/at different time/after different follow up period in the comparison and treatment groups).

Score “YES” if:

- the reported results do not suggest any other sources of bias.

Score “UNCLEAR” if:

- other important threats to validity may be present

Score “NO” if:

- it is clear that these threats to validity are present and not controlled for.

participants, expectations may form an important component of the intervention, so that isolating expectation effects from other mechanisms may be less relevant.

Appendix 3: Critical appraisal of quantitative and qualitative studies included to answer Research Question (2)[32]

1. Is the research aim clearly stated? (Yes/No)

REPORTING:

2. Description of the context? (Yes/No)
3. Description of sampling procedures? (Yes/No)

   - How have the participants been selected, were they the most appropriate?

4. Are sample characteristics sufficiently reported? (sample size, location, and at least one additional characteristic) (Yes/No)
5. Is it clear how the data were collected (eg: *for interviews, is there an indication of how interviews were conducted*? (Yes/No)
6. Methods of recording of data reported? (Yes/No)
7. Methods of analysis explicitly stated? (Yes/No)

METHODOLOGY:

8. Is there a clear link to relevant literature/theoretical framework? (Yes/No)
9. Is the design appropriate to answer the research question? (Yes/No)

   - Has the researcher justified the research design?

10. Was the sampling strategy appropriate to the aims of the research? (Yes/No)

    - Have the researchers explained how the participants were selected?

    - Have the researchers explained why the participants they selected were the most appropriate to provide access to the type of knowledge sought by the study?

    - Have the researchers discussed issues around recruitment? (e.g. why some people chose not to take part)

11. Were the data collected in a way that addressed the research issue? (Yes/No)

    - Were the methods used appropriate and justified?

[32] We developed this tool based on CASP (2006).

- *Did the researcher discuss saturation of data?*

12. Was the data analysis sufficiently rigorous? (Yes/No)

    - *Is there a detailed description of the analysis process?*

    - *Does the data support the findings?*

    - *Is the relationship between the researcher and the participants adequately considered?*

    - *To what extent is contradictory data are taken into account?*

    - *If the findings are based on quantitative analysis of survey data, are multivariate techniques used to control for potential confounding variables?*

13. Has triangulation been applied? (Yes/No)

    - *Data triangulation (location, time and participants)*

    - *Investigator triangulation*

    - *theory triangulation (several theories)*

    - *methodological triangulation*

14. Is the analysis and conclusions clearly presented? (Yes/No)

    - *Have the researchers discussed the credibility of their findings? (e.g. triangulation, respondent validation, more than one analyst)*

    - *Is there adequate discussion of the evidence both for and against the researcher's arguments?*

    - *Are the findings explicit?*

    - *Are the findings discussed in relation to the original research question?*

15. Was there potential for conflict of interest and if so, was this considered and addressed? (Yes/No)

16. Does the paper discuss ethical considerations related to the research? (Yes/No)