# Pre-Analysis Plan

# India Grain Legume Cluster Development

# Meta-information

* This is the fourth version of the Pre-Analysis Plan (PAP) for the Gated Foundation and Aga Khan Foundation Bihar pulses project.
* Updated by IDinsight in June 2018 to add analysis for production and yield outcomes based on post-harvest data collection.
* This PAP was drafted after selection and randomization of program villages, implementation of activities through the Kharif harvest period, digitization of the sampling frame, and piloting of survey instrument, and updated immediately following data collection.
* Clients: Aga Khan Foundation and Bill & Melinda Gates Foundation
* Personnel: Ellen Anderson (Associate), Rupika Singh (Manager), Kate Sturla (Associate Director), Daniel Stein (Chief Economist)

# Background and Models

The Bill and Melinda Gates Foundation, Aga Khan Foundation, and implementing partners are working in 5 districts in Bihar, India to deliver a package of interventions aimed at increasing take-up and productivity of pulses farming in the region. India is a primary consumer of pulses, yet domestic and international production has fallen short of demand, and overall productivity of pulses farming is poor in India. There is evidence that pulses farming improves soil health in addition to being a less water intensive crop.[[1]](#footnote-1) Increased pulses cultivation could also potentially provide rural households a viable source of income, if linked to markets, and/or protein intake, if kept for home consumption.

The project implementation involves a package of interventions aimed at affecting the outcomes of increased production, increased income, and increased household nutrition. These include activities such as seed distribution, extension messaging, best practices training via demonstration plots, formation and support of farmer collectives, and value chain strengthening.

For the 2017 Kharif and 2017-2018 Rabi seasons, IDinsight is evaluating the impact of three different program models developed by AKF and IDinsight – high, medium, and low intensity. The three models vary in the amount of support and inputs provided to farmers. The models are implemented at the village-level. Villages were randomly assigned to one of the three models or control. Randomization was stratified by block. A brief description of the three models is below.

*High Intensity*. The high intensity model consists of demo plots for which inputs and training are provided, free seed distributed at the village-level, and extension messaging and support.

*Medium Intensity.* The medium intensity model consists of free seed distributed at the village-level, and extension messaging and support.

*Low Intensity*. The low intensity model consists of distribution of seed vouchers to farmers which can be redeemed for free seed at a centralized location decided by the Implementing Partner.

Primary Research Question

By how much do each of the three models affect adoption of pulses cultivation?

**Hypothesis**

Farmers will be more likely to adopt pulses cultivation under the high intensity model as compared to the medium intensity; farmers will be more likely to adopt under the medium intensity model as compared to low intensity; and famers will be more likely to adopt under low intensity as compared to control.

**Unit of analysis**

Analysis of cultivation and area outcomes will be on the individual household level.

**Sample**

The sample consists of households who attended an initial general interest meeting about agriculture and pulses cultivation. Sub-group analysis may be conducted.

**Indicators**

* Kharif Cultivation:Binary self-report of whether household planted pulses in Kharif season (yes/no).
* Rabi Cultivation:Binary self-report of whether household planted pulses in Rabi season (yes/no).
* Kharif Area: Self-reported total area of land under pulses cultivation in Kharif season for household.
* Rabi Area: Self-reported total area of land under pulses cultivation in Rabi season for household.

Source

Data on indicators will be collected through household questionnaires administered shortly after the 2017 Kharif season harvest and Rabi season sowing. Self-reported cultivation and area will be collected as part of a crop roster, asking questions related to crops by plot. At the time of the survey, the main pulses crop sown in Kharif, pigeon pea, will not yet have been harvested.

For self-report area under pulses cultivation, respondents will be allowed to answer in any common unit of measurement for that geography including katha, bigha, dhur, acre, hectre, and square meter. The number (decimal option) and unit of measurement (select one option) will be recorded by the enumerator. If answering in a non-standardized local unit (katha, bigha, or dhur), the enumerator will ask the respondent how many of the unit make up one acre, as this varies per district. For intercropped and mixed crops, farmer will be allowed to answer in the ratio of lines or total seed sown. For the purposes of analysis, all responses will be converted to acres.

An initial pilot of the household survey was conducted to asses both validity and reliability of the indicators. An additional pilot was done just prior to data collection.

Validity

As stated in the previous section, data will be collected from surveys on a self-report basis.

Reliability

Back checks will be completed during data collection to assess reliability of the data.

**Analytical Model**

A general linear model will be used to estimate the effect of the treatment on the dependent variables. The model will include dummy variables for block as randomization was stratified by block. The model will be estimated both with and without additional covariates and both results will be presented. Adding covariates will increase the precision of the estimates by reducing standard errors. Standard errors will be clustered at the village level. We will estimate the following regression specification:

, where:

* is the outcome variable for household *i* in village *j*:
  + Kharif cultivation adoption status of household *i* in village *j*
  + Rabi cultivation adoption status of household *i* in village *j*
  + Kharif area under pulses cultivation for household *i* in village *j*
  + Rabi area under pulses cultivation for household *i* in village *j*
* is a binary indicating whether village *j* is in the low intensity treatment
* is a binary indicating whether village *j* is in the medium intensity treatment
* is a binary indicating whether village *j* is in the high intensity treatment
* is a vector of binary block variables
* is a vector of additional household-level covariates:
  + Binary variable for having cultivated pulses at least once in the 2 years previous to the 2017 Kharif season.
  + Binary variable for having pulses crop destroyed by flood
  + Binary variable for binary variables for caste categories General, SC, ST, and OBC
  + Respondent age variable
  + Gender
* is the error term for household *i* in village *j* clustered at the village-level

We will estimate this model in Stata using the following syntax:

areg <depvar> low\_intensity med\_intensity high\_intensity <controls>, a(block) cluster(village)

In order to directly measure the effects one model has versus another, regression post-estimation tests will be performed to determine the equivalence of regression coefficients associated with the treatments.

Tests will be done in Stata using the following syntax:

test low\_intensity =med\_intensity

test low\_intensity = high\_intensity

test med\_intensity = high\_intensity

Although our primary specification considers each treatment arm separately, it is possible that some of the treatment arms will end up having very similar results. Specifically, we believe that the “medium intensity” and “high intensity” treatment arms may end up being implemented similarly and therefore have similar outcomes. If the estimated coefficients for certain treatment arms are very similar, we may choose to pool them in our primary specification to improve power.

# Secondary Research Question 1

By how much does adoption of best practices related to pulses cultivation differ between the three models?

The unit of analysis, sample, and data source are the same as those described in the primary research question section.

**Hypothesis:**

Farmers will be more likely to follow best practices under the high intensity model as compared to the medium intensity; farmers will be more likely to follow best practices under the medium intensity model as compared to low intensity; and famers will be more likely to follow best practices under low intensity as compared to control.

**Indicators**

A package of practices (PoP) summary index will be created for the Kharif season and for the Rabi season. All outcome variables making up the indices will be binary. The outcome variables were chosen based on the respective packages of practices recommended by AKF and implementers in the Kharif and Rabi seasons, and based on which outcomes will be observable at the time of data collection. Summary indices will be constructed using the method from Anderson (2008). Outcome variables will be demeaned and converted to effect sizes by dividing each observation by the variable’s control group standard deviation. The weight of each outcome will be equal to the sum of its row entries in the inverted covariance matrix.

One justification for creating summary indices as opposed to using individual outcomes is that the best practices are bundled together and promoted by the program in a package. Therefore, it makes the most sense to look at adoption of the package as a whole. Other advantages of using summary indices are that one avoids over-testing and reduces random error occurring in the realization of outcomes (Anderson, 2008). We will also perform the analysis on each of the individual indicators in order to understand which specific elements are driving changes in the indices.

Kharif outcome variables for PoP index:

* Binary variables for use of recommended seed treatments: Rhizobium, PSB Culture, Bevestin, Carbendazim, Trichoderma, Thriam, Captan
* Number of times ploughed with recommended range
* Harrowing or planking completed
* Use of cultivator
* Seed priming completed
* Seed rate within recommended range
* Use of line sowing
* Line to line distance within recommended range if line sowing
* Plant to plant distance within recommended range if line sowing
* Binary variables for use of recommended fertilizers: Phosphorous, Nitrogen, Potash, Borax, DAP, Zinc, Sulfur
* Number of weedings within recommended range
* Binary variables for use of recommended weedicide: Pendimythelene, Etrazeen
* Insecticide use or uprooting if stem rot present
* Knowledge of soil moisture during fruiting stage

Rabi outcome variables for PoP index:

* Binary variables for use of recommended seed treatments: Rhizobium, PSB culture, Bevestin, Carbendazim, Trichoderma, Thriam, Captan
* Number of times ploughed with recommended range
* Harrowing or planking completed
* Use of cultivator
* Seed priming completed
* Seed rate within recommended range
* Use of line sowing
* Line to line distance within recommended range if line sowing
* Plant to plant distance within recommended range if line sowing
* Binary variables for use of recommended fertilizers: Phosphorous, Nitrogen, Potash, Borax, DAP, Zinc, Sulfur
* Binary variables for use of recommended weedicide: Pendimythelene, Etrazeen

Validity

This analysis will be purely exploratory. Estimates of the effect of each model on adoption of best practices may be biased as the people who chose to adopt pulses in the each treatment arm and control may be systematically different. For instance, the high intensity arm may contain farmers who adopted pulses cultivation simply because they received free seeds and they will not be motivated to follow best practices, whereas the control group may have more farmers who cultivated pulses because they are intrinsically motivated and will be more likely to follow the best practices

Reliability

Back checks will be completed during data collection to assess reliability of the data.

**Analytical Model**

The analytical model and Stata code will be the identical to that of the primary research question with the exception of the POP index as the dependent variable.

# Secondary Research Question 2

How do pulses production and yields differ between the three models?

**Hypothesis**

Pulses production and yields will be higher under the high intensity model as compared to the medium intensity; production and yields will be higher under the medium intensity model as compared to low intensity; and higher under low intensity as compared to control.

The unit of analysis is the same as described in the primary research question section.

**Sample**

The sample for the production outcome is the same as described in the primary research question section, i.e. households which attended the initial pulses interest meeting.

The sample for the yield outcome is households in the overall sample who reported cultivating pulses for either the Kharif or Rabi seasons during the first round of survey.

**Indicators**

Primary:

Kharif Production: Household’s self-reported total production of Kharif pulses (kgs)

Rabi Production: Household’s self-reported total production of Rabi pulses (kgs)

Secondary:

Kharif Yield: Household’s production per unit of land (kgs/acre) for Kharif pulses from self-reported area and production

Rabi Yield: Household’s production per unit of land (kgs/acre) for Rabi pulses from self-reported area and production

Source

Total pulses production will be collected through a phone surveyed administered following the Rabi 2018 harvest to households who identified growing pulses in either the 2017 Kharif or 2017-2018 Rabi seasons during the round of in-person survey. Shortly after IDisnight’s phone survey, the University of California Davis will be collecting the same production data with a subsample of respondents through an in-person survey. We will also conduct our analysis on this subsample of households who were visited in-person as a robustness check.

Total area under pulses cultivation, as collected in IDinsight’s in-person survey for the primary research question, will be used for the yield calculation. Following questions about pulses production, the phone survey presents the plot and crop areas for pulses plots reported in the earlier in-person survey, and then gives respondents the opportunity to correct the crop areas.

Validity

Primary: The production data is being collected on a self-report basis through phone surveys. There are potential trade-offs with regard to accuracy for using self-report measures as opposed to crop cutting and phone surveys as opposed to in-person surveys. In-person survey data on production collected by UC Davis will be used to understand if our phone survey has introduced any significant biases. If so, we will try to correct for these biases through weighting.

Secondary: The yield analysis will be exploratory. Estimates of the effect of each model on yield may be biased as the people who chose to adopt pulses in each treatment arm and control may be systematically different for the reasons described in the PoP adoption section.

Reliability

Back checks will be completed during data collection to assess reliability of the data.

**Analytical Model**

The analytical model and Stata code will be the identical to that of the primary research question with the exception of the total production and yield as dependent variables.

# Program Exposure

Summary statistics for exposure to aspects of the program listed below will be presented for the sample as a whole, by treatment group, and by district.

* Farmer group membership
* Trainings attended conducted by implementing partner
* Seed received from implementing partner
* Observation or attendance of extension related to pulses cultivation: wall paintings, signboards, mic announcements, public songs, literature distribution, street plays, night meetings, video screenings, puppet shows, and farmer group meetings.
* Knowledge related to line sowing, intercropping, seed treatment, recommended quantity and type of fertilizer, timing of irrigation in accordance with crop stage, and best practices related to disease and pest management for pulses cultivation.

# Cost-Effectiveness Analysis

A costing exercise will be conducted to estimate the total cost of each of the three models at scale. The costs will then be compared with the size of the effects associated with each model. Based on this, a determination will be made on which model is most cost-effective. The cost effectiveness analysis will carried out through a five step methodology described below.

1. Measuring benefits

Respective benefits of the three program models for the purpose of cost effectiveness analysis will be measured through the impact evaluation being carried out by AKF and its implementing partners with support from IDinsight and UC Davis. The cost effectiveness analysis will focus on the outcomes from the primary research question: respective adoption of and area under pulses cultivation for the three models. The outcome of pulses production (and other welfare indicators) may be added following collection of data after the Rabi 2018 harvest.

1. Identifying core activities of the program

IDinsight will compile a list of the core activities of each model which would be replicated at scale. This will be generated based on implementers’ work plans and interviews with AKF and implementers.

1. Understanding marginal program costs, viewpoint, and scale-up

In order to perform a cost effectiveness analysis, we will need to identify the costs which are isolated to the program (marginal costs). These are the costs which would be required to implement the program in other contexts. As state government already have extension infrastructure in place similar to that of the implementing NGOs, it will be important to estimate the differences in cost between the government’s current extension infrastructure and that of the program. We will focus on AKF and implementers’ costs for the three models to date, acknowledging that costs may change at scale.

1. Deciding on a framework

A cost framework will be selected to index costs. The main costing frameworks are ingredients, activities, and COSTAB. The method selected will determine the bounds of potential insights from the analysis.

1. Comparing the three program models

Based on the first four steps, the cost effectiveness of the three program models will be calculated and compared.

Work Cited

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A

reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American*

*statistical Association*, *103*(484), 1481-1495.

1. Bill and Melinda Gate Foundation and Aga Khan Foundation Project Proposal. [↑](#footnote-ref-1)