Kartu Prakerja Impact Evaluation Analysis Plan
September 2021

Research team: Vivi Alatas, Rema Hanna, Benjamin Olken,
Achmad Maulana, Elan Satriawan, and Sudarno Sumarto

Primary Outcomes

The six primary outcomes we will examine are:

- Labor force outcomes (employed; new job or business since program application; wage; monthly household income, hours worked; job search and activity preparing a new business; job satisfaction, confidence in business skills, child care responsibilities, sectoral change subject to data availability)
- Consumption smoothing (asset sales, loans, transfers, migration, and self-reported and subjective consumption)
- Psychometric outcomes (depression; self-efficacy)
- Digital skills and comfort (use of internet in job, comfort with and usage of e-money, preferences for e-money vs. phone credits as survey compensation, use of platforms)
- Types of trainings chosen (courses selected, amount of training budget spent).
- Approval of government COVID response and preferences about government programs

We will also examine 'first stage' outcomes that measure program usage (e.g., program uptake, obtaining training certificates).

Heterogeneity Analysis: Prime dimensions of heterogeneity to examine

- *Multiple applications*: e.g., for households that apply in batch X, compare impacts for winners from batch X with impacts from those who lose batch X but reapply and win in a subsequent batch. We will focus on batches for which there is immediate and/or short delay between the announcement of one batch and the application to the next batch.
- *Gender*: for households where multiple family members apply in the same batch, how do outcomes differ if a male vs. female applicant is randomly selected as the winner in that batch. We can also examine heterogeneity in impacts by gender of respondent.

Note that for these outcomes, we will consider the primary outcomes above, as well as additional primary outcomes from the administrative data (e.g., number of reapplications, type of training course chosen, number of training courses completed, share of available training budget spend, and time to complete first training).

When possible (i.e. depending on timing of data source), we will do heterogeneity based on whether those randomized to receive the program are currently / recently getting the stipend as of the time of the survey.

We will also do secondary heterogeneity analysis based on demographics (age, education, gender, rural/urban, java/off java, and baseline occupation if available).

Descriptive Analysis of Program Uptake

We will also provide descriptive analysis of program uptake using baseline data matched to administrative data for eligible populations, including marginal value of income (e.g. consumption, wages), baseline internet access (e.g. cell phone coverage, smartphone/laptop ownership), gender / recent maternity status, previous employment, disability, and other demographics. This analysis will also allow us to assess the degree to which PraKerja helped fill exclusion gaps from other GoI social assistance programs.

Data

We will analyze the following datasets for outcome analysis, matched to the Kartu Prakerja administrative data:

- Sakernas August 2020
- Susenas September 2020
- Sakernas February 2021
- Susenas March 2021
- Sakernas August 2021
- Susenas September 2021
- Survey conducted by the research team in cooperation with KP management starting August 2021

In addition, for the descriptive analysis of program update, we will examine baseline (e.g. 2018-2020) Sakernas and Susenas data, as available.

We focus our analysis on all Prakerja batches conduced prior to the outcome data, except batches 1 and 15.

Regression specification

A given individual's probability of winning depends on a series of input variables and the provincial quota. The primary impact evaluation regressions specifications are as follows:

1. Reduced form, controlling for all input variables used to construct the randomization weights

$$Outcome_{ib} = WinBatch_{ib} + Province_{ib} \times Inputvars_{ib} + X_i + \epsilon_{ib}$$

where X_i are individual level baseline controls from the administrative data, chosen via LASSO.

2. Individual IV, controlling for all input variables used to construct the randomization weights

$$Outcome_{ib} = EverWin_i + Province_{ib} \times Inputvars_{ib} + X_i + \epsilon_{ib}$$

where $WinBatch_{ib}$ is used to instrument for $EverWin_i$, which captures if individual i ever received the program prior to the date of the survey.

3. Family IV, controlling for all input variables used to construct the randomization weights

$$Outcome_{ib} = FamilyEverWin_i + Province_{ib} \times Inputvars_{ib} + X_i + \epsilon_{ib}$$

where $WinBatch_{ib}$ is used to instrument for $FamilyEverWin_i$, which captures whether an individual from family i ever wins the program prior to the date of the survey.

All standard errors are clustered by family. We report robust clustered standard errors, and randomization-inference-based p-values.

Regressions are run batch-by-batch, as well as stacked and pooled. For stacked and pooled regressions, we cluster for observations that appear in multiple batches.

We will also run versions above where we control for $Province_{ib} \times RandWeight_{ib}$ where $RandWeight_{ib}$ is the PMO's calculation of the randomization weight used:

1. Reduced form, controlling for all input variables used to construct the randomization weights

$$Outcome_{ib} = WinBatch_{ib} + Province_{ib} \times RandWeight_{ib} + X_i + \epsilon_{ib}$$

where X_i are individual level baseline controls from the administrative data, chosen via LASSO.

2. Individual IV, controlling for all input variables used to construct the randomization weights

$$Outcome_{ib} = EverWin_i + Province_{ib} \times RandWeight_{ib} + X_i + \epsilon_{ib}$$

where $WinBatch_{ib}$ is used to instrument for $EverWin_i$, which captures if individual i ever received the program.

3. Family IV, controlling for all input variables used to construct the randomization weights

$$Outcome_{ib} = FamilyEverWin_i + Province_{ib} \times RandWeight_{ib} + \mathrm{X_i} + \epsilon_{ib}$$

For batches i where the first stage regression of $EverWin$ on $WinBatch$ is 0.75 or lower, and pooled analysis including such batches, we can also run additional specifications to increase power in the spirit of Rotnitzky and Robins (1995) for missing data but generalized for cases when subsequent first stage is due to observed randomziation: we will drop all households who lose batch i, enter and win batch $j > i$, and instead will correspondingly increase the weight on comparable households who lose batch i , have the same subsequent application history, but lose batch j.

Note that the survey was sampled by batch. For people who are in the survey in batch *i*, but were sampled in batch *j>i*, we will reweight to make sure that the sample is balanced based on original status in batch sampled.

Balance analysis

We will examine baseline balance batch-by-batch using the above regression equations as follows, using administrative data and available matched waves of government-run Sakernas and Susenas household surveys:

- Administrative Data: city (kotamadya), gender, did another family member apply

- Sakernas: Gender, # members in HH, employed in the last week, use internet at work, had a course with a certificate, hours worked last week, city, born and live in same regions
- Susenas: Gender, # members in HH, employed in the last week, use internet at work, had a course with certificate, and hours worked in the last week