Worker Paid Leave Usage Simulation (Worker PLUS) Model
Issue Brief: A Guide to Performing Example Simulations

OVERVIEW

This brief provides a step-by-step guide to performing example simulations using the Worker Paid Leave Usage Simulation (Worker PLUS) model developed by IMPAQ International and the Institute for Women's Policy Research (IWPR) for the Chief Evaluation Office at the U.S. Department of Labor (DOL). With this guide, users should be able to replicate the provided example of model running using either the Python or the R simulation engine, and to check how the simulation results compare against actual program data for existing state programs in California, New Jersey, and Rhode Island. To best facilitate the replication exercise by reducing runtime, we choose Rhode Island as the example state, for which the American Community Survey (ACS) Public Use Microdata Sample (PUMS) is a lightweight dataset with a sample size of 17,475 eligible workers. The replication of results for other states follows the same steps, but with a longer runtime.

We first provide guidance on how to prepare the model for running, including confirming hardware and software environments, selecting correct datasets, launching the graphical use interface (GUI), and selecting state population and program parameters. Next, we show how the simulation should be configured for comparison between simulation results and actual program data, and for replication of results by specifying simulation method and random seed control. We then discuss the simulation results and compare them against Rhode Island program data. In concluding remarks, we provide additional details on how this exercise can be generalized for different states.

SETTING UP THE MODEL

The model has been tested on mainstream Windows machines. The model can be launched via a single Python command and offers a user-friendly interface for easy configuration.
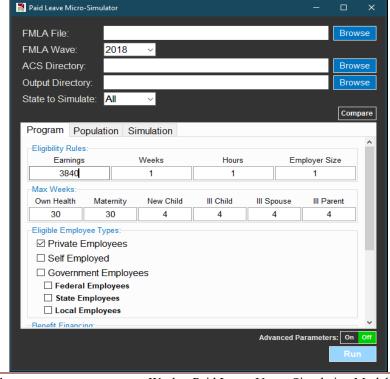
The model has been tested on mainstream workplace and home computers with Intel i5 and i7 multicore processors. Minimum RAM needed is 8 GB, although we recommend 16 GB of RAM or more for better performance. For disk space requirements, the file size for model code is 177 MB,

To facilitate understanding of the potential impacts of different policy alternatives on workers' leave-taking behaviors and program costs, the U.S. Department of Labor's Chief Evaluation Office contracted with IMPAQ International, and its partner Institute for Women's Policy Research (IWPR), to develop the Worker Paid Leave Usage Simulation (Worker PLUS) model, an open-sourced microsimulation tool based on public microdata and predictive modeling. The model and other relevant materials are publicly available at [*hyperlink*].

In this issue brief, we provide a step-by-step guide to performing simulations using the Worker PLUS model developed by IMPAQ International and the Institute for Women's Policy Research for the Chief Evaluation Office at the U.S. Department of Labor.

We use the 2018 U.S. Department of Labor Family and Medical Leave Act Employee Survey and 2014–2018 American Community Survey Public Use Microdata Sample for the state of Rhode Island to perform an example simulation. We then use actual data from the Rhode Island program from 2014 to 2018 to benchmark the simulation results. With this guide, users can replicate model launch, configuration, and execution.

Exhibit 1: Graphical User Interface after Model Launching

and the file size of all input data except ACS PUMS is 81 MB. ACS PUMS is the largest set of model files by file size. Users can choose one of two options to download the ACS PUMS data along with the model.[i] In the **ACS File Download Option A**, the ACS PUMS is available only for the default setting (ACS Year = 2018, and State of Work = True), with a file size of 11.4 GB. In **ACS File Download Option B**, all ACS settings of the model are enabled (ACS Year = 2016, 2017, or 2018, and State of Work = True or False) with a larger database, with a file size of 81.5 GB. The additional ACS files available from the **ACS File Download Option B** allow users to perform simulations for evaluating programs using historical data, or exploring differences in simulation outcomes using different definition of workers' states. The **ACS File Download Option A** is recommended for saving local disk space if users do not need to perform these simulations. The model offers two simulation engines via Python and R, two popular open-source programming languages. The minimum versions required to run the model are Python 3.7 for the Python engine, and R 4.0.0 for the R engine. The model is developed and tested on Windows operating systems. The minimum version required is Windows 10.[ii]

Both Python and R simulation engines are ported to a GUI which is coded in Python *tkinter*. To launch the GUI, in a Windows Command Prompt or PowerShell, change the directory to the project folder, which contains the file *Microsimulator.py*, and execute:
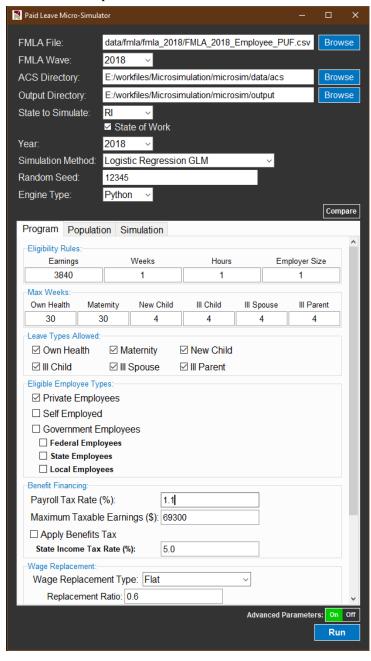
```
python Microsimulator.py
```

A GUI, as shown in **Exhibit 1**, will then be launched.

RUNNING THE MODEL
Users can navigate to any desired local directory where Family and Medical Leave Act (FMLA) and ACS PUMS data are stored but should ensure that the ACS directory follows the required sub-directory structure, and the data files within the directory follow the required file naming format. For states with sufficient program data published, including Rhode Island, users can set Existing State Program, which will automatically fill all Program and Population parameters for the user.

Exhibit 2: Graphical User Interface with Advanced Parameters

To perform the simulation illustrated in this brief, the user should first turn on the **Advanced Parameters** option at the right bottom corner of the GUI (above the *Run* button) to show all model parameters and options, as in **Exhibit 2**, and then apply the following configuration in the GUI:

- **FMLA File**: This is the file path of the 2018 DOL FMLA Employee Survey data file, which must be in CSV format. The correct file path can be verified by ensuring that path ends with the ".csv" extension.[iii] The user can use the Browse button on the right to navigate to any local directory where the FMLA data file is located.
- **FMLA Wave**: This is the wave year of the FMLA Employee Survey data. The user can choose from 2012 or 2018, the two years for which FMLA data are used for modeling. The user should set this field to 2018 to perform the simulation described in this brief.

- **ACS Directory**: This is the directory of the ACS PUMS data. As with the FMLA File field, the user can use the Browse button to navigate to the desired local directory. While the ACS Directory itself can be any directory on the local machine, the model requires that the sub-directories of the ACS Directory follow a specific structure. Assuming that the user has chosen **ACS File Download Option B**,[iv] and the ACS Directory is *./acs*, the model then requires the following sub-directory structure (using ACS PUMS 2014–2018 as an example):
    - *./acs/2018/household_files*
    - *./acs/2018/person_files*
    - *./acs/2018/pow_household_files*
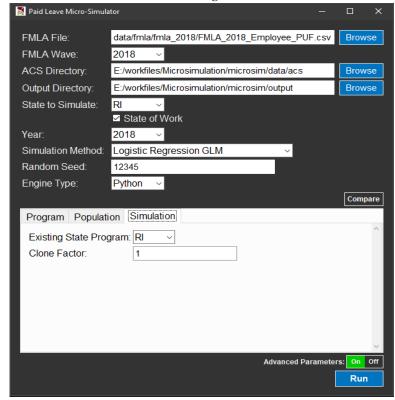   - *./acs/2018/pow_person_file*

  where the former two sub-directories respectively store the original state-level ACS PUMS household and person files for the fifty states plus the District of Columbia, and the latter two store the repartitioned state-level ACS PUMS data files based on the ACS PUMS place-of-work variable POW. Users should not rename the data files under these directories, as files with an incorrect file name will not be recognized by the model code.

- **Output Directory**: This is the directory where model output files will be saved. User can use the Browse button to select any desired local directory for model output.

- **State to Simulate**: This field should be set to RI from the dropdown list so that the model will use the Rhode Island ACS PUMS.

- **State of Work**: This box should be checked so the place-of-work (POW) versions of the state of Rhode Island ACS PUMS data files are used. These data files are as follows:
    - *./acs/2018/pow_household_files/h44_ri_pow.csv*
    - *./acs/2018/pow_person_files/p44_ri_pow.csv*

- **Year**: This field should be set to 2018, the ending year of ACS PUMS 2014–2018.

- **Simulation Method**: To replicate results in this brief, this field should be set to *Logistic Regression GLM*, representing the logistic model built from the generalized linear model (GLM) family.[v] This is the default simulation method for both Python and R, and it has been verified to perform estimation using the same solver algorithm in both languages.

- **Random Seed**: The seed value can be any integer between 0 and 2^{32}. The simulation results will remain identical on the same operating system as long as the same seed value is specified. We recommend using the same integer for any comparison-based analysis for easy tracking and replication. The results in this brief can be replicated using the random seed value 12345 on a Windows 10 Pro operating system with OS Build version 19041.630. The results may slightly vary for different Windows operating systems and OS Build versions.

- **Engine Type**: To replicate results in this brief, Python should be selected.

- The user does not need to enter any parameters in the **Program** and **Population** tabs, as they will all be automatically filled in once **Existing State Program** is set to "RI" under the **Simulation** tab, as shown in **Exhibit 3**. In particular, the **Program** parameters will be filled in with parameters of the Rhode Island program, and the **Take-Up Rates** under the **Population** tab will be filled in with rates estimated from Rhode Island program statistics. These parameters are

Exhibit 3: Specifying Program and Population Parameters via Existing State Program

estimated from the average of Rhode Island program data during 2014–2018, corresponding to the same five-year period of the ACS microdata, the target sample used for simulation.

- The user should leave **Clone Factor** as 1 under the **Simulation** tab.[vi] If this integer factor is greater than 1, then the ACS sample will be cloned with population weight *PWGTP* scaled down proportionally, allowing user to obtain more granular simulation results for more ACS persons while leaving the sum of population weights across persons unchanged. In principle, the clone factor can be any positive integer value. However, we recommend values no larger than 2, which would lead to a sample size of over 17,000 eligible workers even for the smallest ACS PUMS state sample (Wyoming) under eligibility rules of existing programs (e.g., Rhode Island). For the Rhode Island ACS PUMS state sample, we found that a **Clone Factor** of 1 led to 17,475 eligible workers, which should be sufficient for estimating state-level statistics such as total program benefit outlays.

- The **Compare** button shown on top of the **Program**, **Population**, and **Simulation** tabs is for users to add comparison simulations using the same settings in the main GUI panel fields (i.e., **FMLA File**, **FMLA Wave**, **Year**, **Simulation Method**, etc.) but different settings under the **Program**, **Population**, and **Simulation** tabs. This brief focuses on executing a single simulation; therefore, users should not press the **Compare** button to replicate the results.[vii]

- After the configuration above, the user can click the **Run** button to execute the model for a given choice of **Engine Type** (Python or R). For Rhode Island, we expect a total runtime per simulation of about 2 to 5 minutes, depending on the processor speed and RAM of the machine.

SIMULATION RESULTS

The model can produce results for program outlays and program case counts that closely track actual program data for most leave reasons. Total benefit outlay is underestimated by about 11.7%. This is mainly caused by underestimation of leave takers due to maternity since the simulation method Logistic Regression GLM predicts fewer leave takers than other simulation methods.

After the simulation has been completed, the GUI will present a result window which shows a summary of program benefit outlays, including the total outlay and the outlay for each of six paid leave types, as shown in **Exhibit 4**. For brevity of this brief, we focus on the results shown under the **Summary** tab of this window. The results under the **Benefit Financing** and **Population Analysis** tabs are discussed in more detail in two separate issue briefs in this series.[viii]
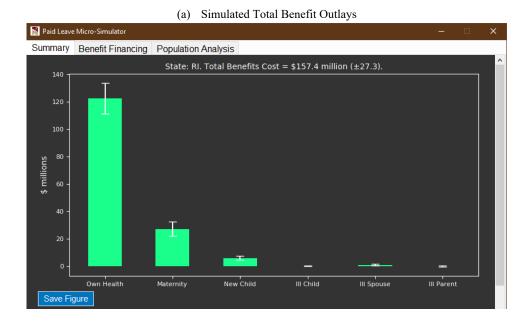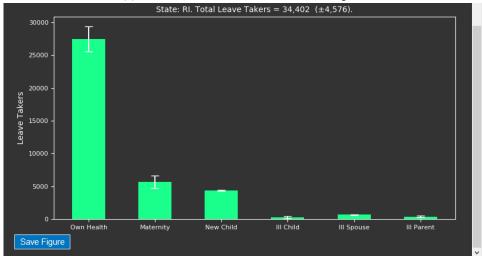
Exhibit 4: Simulation Result Window

(a)   Simulated Total Benefit Outlays

(b) Simulated Total Leave Takers under Program

Note: Simulation results are based on eligible workers in Rhode Island identified from the 2014–2018 ACS PUMS state samples. The sample size of eligible workers is 17,475, representing an underlying worker population estimate of 372,082. Each bar represents a leave reason: *Own Health* = own illness of worker; *Maternity* = maternity disability of worker; *New Child* = bonding with newborn or newly adopted child; *Ill Child* = caring for an ill child; *Ill Spouse* = caring for an ill spouse; and *Ill Parent* = caring for an ill parent. For each bar in (a) and (b), the 95% confidence interval derived from the ACS replication weights is also displayed.

We then compare the estimates above against program data downloaded from the Rhode Island program website.[ix] The outlay estimates shown in **Exhibit 4** (a) are stored as a CSV file in the folder *output_[datestamp]_[timestamp]_main simulation* under the *Output Directory* specified by the user in the GUI, named *program_cost_ri_[datestamp]_[timestamp].csv*. In the same folder, the estimates of leave takers under the program shown in **Exhibit 4** (b) are stored in the file *program_progtaker_ri_[datestamp]_[timestamp].csv*. In the output folder and file names, *[datestamp]* and *[timestamp]* are respectively the date stamp and time stamp of the time when the **Run** button was pressed to execute the simulation, thus they uniquely identify the output folder and files.

Exhibit 5 reports two sets of comparisons of program statistics between actual program data and simulated values, one for benefit outlays, and the other for number of paid cases. Overall, most of the simulated program outlays and paid case counts can closely track the actual program data for most leave reasons. The total program benefit outlay is $157.4 million, 11.7% lower than the actual program outlay of $178.2 million. Further investigation into the simulated values for leave taker counts under the program indicates that this underestimation is mainly driven by the underestimation in leave takers due to the leave reason *Maternity*. Our model testing results reported in a separate brief have shown that the underestimation in leave taker counts originates from the choice of simulation methods.[x] In particular, the *Logistic Regression GLM* method chosen in this brief is one that typically simulates fewer leave takers than other methods. The underestimation would be mitigated when other simulation methods are chosen for the field **Simulation Method** shown in **Exhibit 2**, such as *Ridge, K Nearest Neighbor, Naïve Bayes, Support Vector Machine, Random Forest,* and *XGBoost*. For example, the number of paid cases for leave reason *Maternity* is simulated to be 5,657 under *Logistic Regression GLM*, while it would be 9,463 under *Random Forest*, fairly close to the actual count 9,444.

Besides the leave taker count under program for *Maternity*, we also note that the outlays are underestimated by 24% for family caring reasons, including *New Child, Ill Child, Ill Spouse,* and *Ill Parent,* as shown in the second row of **Exhibit 5** under Total Program Benefit Outlay. The leave taker counts for these reasons, however, are not underestimated, as shown in the paid cases rows. This suggests potential underestimation in family care leave lengths, arguably the most unpredictable set of outcomes that can be affected by many unobserved factors, such as health status, time preferences, and trade-off between family care duty and job security.

Exhibit 5: Comparison between Actual and Simulated Program Statistics, Rhode Island, 2014–2018

Description of Program Statistic	Simulated Value	Actual Program Data
Total Program Benefit Outlay	$157.4 million	$178.2 million
- Program Benefit Outlays, Own Health and Maternity	$149.7 million	$167.4 million
- Program Benefit Outlays, New Child, Ill Child, Ill Spouse, and Ill Parent	$7.7 million	$10.1 million
Paid Cases	38,751	43,177
- Own Health	27,473	28,333
- Maternity	5,657	9,444
- New Child	4,355	4,338
- Ill Child	240	199
- Ill Spouse	648	633
- Ill Parent	378	355

Note: Simulated values are produced by the Worker PLUS model. Simulation results are based on eligible workers in Rhode Island identified from the 2014–2018 ACS PUMS state samples. The sample size of eligible workers is 17,475, representing an underlying worker population estimate of 372,082. Historical program data are five-year averages between 2014 and 2018 obtained from https://dlt.ri.gov/lmi/datacenter/uitdi.php.

The above example simulation for Rhode Island workers can be easily generalized to California and New Jersey, the other two states for which we have built into the model the program rules and take-up rates using historical program data from 2014 to 2018. All steps remain the same to implement the model for these states, except that the **Existing State Program** under the **Simulation** tab and the **State to Simulate** in the main GUI panel need to be specified as "CA" or "NJ."

CONCLUDING REMARKS

- This brief has provided a step-by-step guide on how to use the Worker PLUS model to perform an example simulation using parameters of an existing state program (Rhode Island) and obtain simulation results that can closely track actual program data. By following the guidance provided, model users should be able to replicate the simulation and comparison results.
- The Worker PLUS model provides a framework where users can configure many features of the program, underlying population, and simulation methods. To produce simulation results that can closely track program data of the other available states (i.e., California and New Jersey), we have coded the model so that parameters under the **Program** and **Population** tabs in the GUI can be automatically filled in based on the user's choice of **Existing State Program** under the **Simulation** tab. The open-source nature of the model will facilitate future development of this feature by adding in data on state program rules and program take-up rates as program data accumulate across more states.
- For easy reference, a summary spreadsheet of historical program data and an archive of the underlying source program documents and reports will be made available to users.[xi] The continuous monitoring and archiving of these source statistics should be a crucial component of model development, validation, and use, in order to ensure proper benchmarking of program outlays and calibration of program take-up.

---

[i] The ACS File Download Options A and B are described in more details in IMPAQ (2021). Worker Paid Leave Usage Simulation Model User Manual.

[ii] For additional details of hardware and software requirements, see IMPAQ (2021). Worker Paid Leave Usage Simulation Model User Manual.

[iii] The 2018 DOL FMLA Employee Survey data files are available at https://www.dol.gov/agencies/oasp/evaluation/fmla2018. The data file is available in DTA (for Stata) and SAS7BDAT (for SAS) formats. After downloading the data, users should first convert the data file into CSV format, which can be performed in both Stata and SAS respectively.

[iv] A similar sub-directory structure applies for ACS PUMS data files downloaded via the **ACS File Download Option A**, in which case there is only one folder named "2018" to be placed under the ACS directory specified by the user.

[v] In contrast, the Logistic Regression Regularized option in the **Simulation Method** dropdown menu represents the logistic model with a regularization term in the objective function to account for model complexity.

[vi] More details on how **Clone Factor** affects the ACS population weight variable PWGTP is provided in IMPAQ (2021). Worker Paid Leave Usage Simulation Model User Manual.

[vii] For additional details on comparing simulations using the *Compare* button, see Worker Paid Leave Usage Simulation Model User Manual (2021).

[viii] The details of results under the Benefit Financing tab are reported in IMPAQ (2021). Worker Paid Leave Usage Simulation (Worker PLUS) Model. Issue Brief: Benchmarking Results of the Benefit Financing Module's Payroll Tax Revenue Estimates. The details of results under the Population Analysis tab are reported in IMPAQ (2021). Worker Paid Leave Usage Simulation (Worker PLUS) Model. Issue Brief: A Guide to Performing a Policy Simulation of a Parental Leave Program for Federal Civilian Employees.

[ix] Rhode Island program data are available from https://dlt.ri.gov/lmi/datacenter/uitdi.php. Program data for other states can be accessed from https://edd.ca.gov/about_edd/Quick_Statistics.htm (California) and https://myleavebenefits.nj.gov/about/stats (New Jersey).

[x] The details of predictive performance of different classifiers are reported in IMPAQ (2021). Worker Paid Leave Usage Simulation (PLUS) Model. Issue Brief: Model Testing.

[xi] The spreadsheet and the archive are available from the *state_program_statistics* folder, in the same directory as the file *Microsimulator.py*.