This document describes the replication package for "What Do Data on Millions of U.S. Workers Reveal about Life-Cycle Earnings Risk?" (Guvenen, Karahan, Ozkan and Song 2021).

We used Stata and Matlab programs to generate the empirical results from the Social Security Administration (SSA)'s the Master Earnings File (MFE). The SSA does not allow us to share the micro data or the intermediate files obtained using it. To facilitate replication for researchers with access to the Master Earnings File at the SSA, we provide all the STATA programs that process the micro data and the Matlab programs that produce the figures in the paper. We now explain them in detail.

There are four subfolders in this replication package:

1. **STATA programs:** According to our confidentiality agreement with the SSA, we are not allowed to share the micro data or any of the interim datasets constructed as part of this project. To facilitate replication for researchers with access to the MEF, we provide all the STATA programs that process the micro data in this "STATA programs" subfolder. These programs are also available on the SSA servers.

    - The user first needs to run Gen_Labor_Sample2013.do to generate the samples used in our analysis as we describe in Section 2 of the paper.
    - Gen_Labor_Sample_3EIN.do is a similar program that generates the sample with employer identification numbers used in our stayer-switcher analysis in Section 3.5.
    - Merge_DIB.do (and DIB_AgeDum.do) merges disability benefits to our main sample, which is used in Section 3.6.
    - SE_imputation_Qt_LN.do generates the quantile regression coefficients we estimated to impute self-employment income above the SSA taxable limit. We explain our imputation procedure in Appendix A.1. Furthermore, LifeCycleProfile_LABOR_imputation.do generates the sample between ages 22 and 60, which is used in this analysis.

    For other do files under the "STATA programs" subfolder see the excel file "GKOS_2021_replication.xlsx", which links the output obtained from the MFE that is used in each figure in the paper to the STATA code that generates this output. These intermediate files from the MFE are then used in our Matlab programs to plot the figures in the paper.

2. **Matlab plotters:** This subfolder includes all of the Matlab programs that produce the figures in the paper. These programs use intermediate files produced by the Stata programs using the MEF as explained above.

    The excel file "GKOS_2021_replication.xlsx" provides a mapping between the figures in the paper and the Matlab programs that have been used to generate them.

    We also include the custom Matlab packages (export_fig, line_fewer_markers_v4, subaxis) that the Matlab programs use.

3. **PSID programs:** Section 3.6 uses data from the Panel Study of Income Dynamics (PSID). The folder "PSID programs" contains the (publicly available) raw PSID data as well as the STATA programs that process these data and generate Tables 2 and 3 in the manuscript. The user can find a detailed readme file (PSID_readme.pdf) that explains how one can replicate our results from the PSID.

4. **Estimation:** In this subfolder we provide the files that are necessary to estimate the benchmark income process we present in Section 6 of the paper. These files include Fortran source codes, a makefile as well as the input files for the moments targeted in the estimation. The Fortran code can be modified to estimate the other processes considered in Section 6 and Appendix D. We also provide a detailed readme file under the same folder, which explains each file in the folder and how one can run this code.

The spreadsheet **GKOS_2021_replication.xlsx** contains the data for each figure in the paper and provides a mapping on which Stata and Matlab programs have been used to generate them. To produce any of the figures, one should first run the corresponding Stata program on the MEF and then run the Matlab program using the output of the Stata program.