

README – Replication Package

Paper: Click, Code, Earn: The Returns to Digital Skills

Authors: Antonio Soares Martins Neto, Yan Liu, Saloni Khurana, and Juan Porras

Overview

This replication package provides the Stata codes used to reproduce the results of the paper Click, Code, Earn: The Returns to Digital Skills.

Dependencies

The analysis was conducted using Stata 18. The following Stata packages must be installed prior to running the code:

- `ssc install winsor2`
- `ssc install estout` (for regression tables, if used)
- `ssc install listtab`

Folder Structure

The main folders included in the package must be in the same directory for the replication to run correctly:

```
Input
├── Clean_data_v4.dta
├── v21_lightcast_original_postings_by_country_year.dta
├── v21_lightcast_postings_by_country_2021_2024.dta
├── v21_lightcast_postings_by_occupation_2021_2024.dta
├── v21_lightcast_postings_by_sector_2021_2024.dta
├── v21_lightcast_postings_by_isco_and_naics_2021_2024.dta
├── v21_sample_country_trends.dta
├── itintensity_naics2_cou.dta
├── ILO_labor_force_data.dta
├── inputs raw postings processing
│   ├── Lightcast_digital_skills_mapping_rds.csv
│   ├── Lightcast_soft_skills_mapping_rds.csv
│   ├── ai_skills_mapping_03.csv
│   ├── v21_isco_title_names_by_top_frequencies
│   ├── v21_non_null_data_2_dataset_2021_2024_traditional_ai_v4.csv
│   └── wb_income_fy26_25_comparison.dta
├── inputs it intensity
│   ├── icio_build.do
│   ├── sectors.dta
│   └── tiva_naics2.dta
└── Program
    ├── 0_Clean_data.do
    ├── 1_Descriptive_stats.do
    ├── 2_Estimates.do
    ├── Master.do
    ├── Returns to Digital Skills - Lightcast RDS dataset construction.html*
    └── Returns to Digital Skills - Skills Wordclouds.html*
Output
├── Figures
└── Tables
```

How to Run

1. Update main directory path in the master file Master.do.
2. Run the Master do-file. This will execute the following in sequence:
 - 0_Clean_data.do → Cleans and harmonizes the dataset
 - 1_Descriptive_stats.do → Produces descriptive statistics and summary figures
 - 2_Estimates.do → Generates econometric estimations and final figures/tables
3. All outputs (figures and tables) will be saved in the Output/ folder.

Code Description

STATA

- [0_Clean_data.do](#) – Cleans and harmonizes the input dataset generated from the Lightcast postings processed in Databricks. It standardizes occupation and sector codes (ISCO-08, NAICS2), harmonizes education and employment variables, applies winsorization to wage data, and produces the cleaned analytical dataset used across all tables and figures.
- [1_Descriptive_stats.do](#) – Produces descriptive statistics on skill distributions, education, and occupations; creates baseline tables and figures.
- [2_Estimates.do](#) – Runs regression models and generates the estimation results used in the paper.
- [IT_intensity.do](#) – Constructs and merges the IT-intensity indicators derived from OECD TiVA/ICIO data (via itintensity_naics2_cou.dta) into the analytical dataset. It computes the interaction terms between sectoral IT intensity and digital-skill requirements used in the robustness checks and heterogeneity analyses.

PYTHON

- [Returns to Digital Skills - Lightcast RDS dataset construction](#)
Core data-engineering notebook used to construct the Returns to Digital Skills (RDS) dataset from the Lightcast Global Job Postings. It performs large-scale cleaning, skill extraction, digital-skill classification in PySpark. The resulting harmonized datasets are exported to csv format for further cleaning and analysis in Stata.
- [Returns to digital skills - Skills Wordclouds.html](#)
Generates the digital-skill word clouds for the countries included in the analytical sample of the paper “Click, Code, Earn: The Returns to Digital Skills.” Using the cleaned Lightcast dataset, it visualizes the most frequently mentioned digital, intermediate, and advanced skills by skill category and income group. This notebook reproduces the descriptive visualizations corresponding to the paper’s figures and was executed within the Databricks environment.

Data Availability Statement

This replication package draws on Lightcast job-posting data integrated with international occupational and sectoral classifications, supplemented by auxiliary datasets from the ILO and industry IT-intensity measures. All input datasets must be in the Input/ folder prior to running the master do-file. The main datasets were obtained from Lightcast under the World Bank licensing agreement and were last accessed in June 2025.

1. Lightcast Global job postings

Description: The Lightcast Global Job Postings Database is a proprietary dataset containing worldwide online job postings collected from several sources. It includes detailed metadata on occupations, industries, skills, employers, wages, and posting timelines. All datasets listed below are filtered, aggregated, or harmonized derivatives created using Databricks (PySpark) and Stata processing scripts.

Citation: Lightcast (2025). *Lightcast Global Job Postings Database* [Proprietary dataset]. Accessed June–July 2025 under institutional license.

Availability: Restricted. Access requires a commercial data agreement with Lightcast.

Derived Datasets (Version 21, filenames preserved)

- v21_non_null_data_2_dataset_2021_2024_traditional_ai_v4.csv
- Clean_data_v4.dta
- v21_lightcast_original_postings_by_country_year.dta
- v21_lightcast_postings_by_country_2021_2024.dta
- v21_lightcast_postings_by_sector_2021_2024.dta
- v21_lightcast_postings_by_occupation_2021_2024.dta
- v21_lightcast_postings_by_isco_and_naics_2021_2024.dta
- v21_sample_country_trends.dta
- v21_isco_title_names_by_top_frequencies.csv
- v21_developed_countries_ds_skills_by_frequencies_non_null_sample.dta
- v21_developing_countries_ds_skills_by_frequencies_non_null_sample.dta
- Lightcast_digital_skills_mapping_rds.csv

Manually curated mapping of Lightcast skill IDs to digital-skill levels, developed by the research team. Each skill code is classified as Basic, Standard, or Advanced according to the taxonomy defined in Click, Code, Earn: The Returns to Digital Skills.

- Lightcast_soft_skills_mapping_rds.csv

Mapping of Lightcast skill IDs associated with top soft skills, including cognitive, interpersonal, management, and personal attributes.

- ai_skills_mapping_03.csv

Manually curated list of skills identified as Artificial Intelligence (AI) or Generative AI (GenAI) related. The mapping distinguishes between traditional AI and GenAI development/literacy skills, following the definitions used in the paper.

2. IT-Intensity Dataset — sectors.dta

Description: This sectoral dataset serves as an input dataset containing sector-level value-added components from the OECD Inter-Country Input-Output (ICIO) and Trade in Value-Added (TiVA) databases. These data are combined with World Bank sectoral aggregations and harmonized to NAICS 2-digit industries using an author-developed NAICS–TiVA concordance (tiva_naics2.dta). The harmonized sectoral values are then used to compute IT-intensity ratios by country and industry, producing the country–industry IT-intensity measures used in the empirical analysis.

Citation

- **OECD.** (2023). *OECD Inter-Country Input-Output (ICIO) Tables and Trade in Value-Added (TiVA) Database* [Data sets]. OECD. <https://data-explorer.oecd.org/>
- **Boy, H. C. (2023).** sectors.dta [Author-constructed dataset]. World Bank Group. *Available in the replication package.*
- **World Bank Group.** (2024). tiva_naics2.dta [Author-constructed dataset]. *Available in the replication package.*

Modifications: IT intensity is calculated for each country and industry using 2020 data from the Trade in Value-Added (TiVA) dataset. Because IT intensity tends to evolve slowly, reflecting structural characteristics of industries, relying on pre-period data helps address concerns about simultaneity with our 2021–2024 study period.

Derived Datasets:

- itintensity_naics2_cou.dta

3. World Bank Income Classification — wb_income_fy26_25_comparison.dta

Description: This dataset provides a country-level comparison of the World Bank income classification between fiscal years 2025 and 2026. It is constructed by the authors using publicly available World Bank income classification data and maps changes in country income groups across the two fiscal years.

Citation

World Bank Open Data Blog: *Understanding country income — World Bank Group income classification*. Publicly available from the World Bank in <https://blogs.worldbank.org/en/opendata/>

4. ILO Labor-Force Data — ILO_labor_force_data.dta

Description: This dataset contains annual labor-force indicators (employment, labor-force participation, working-age population 15+) sourced from the ILOSTAT database. It includes all available countries, both sexes, and age group 15+. Extracted and standardized as an input to compute labor-force denominators.

Citation: International Labor Organization. (2025). ILOSTAT: Labor force by sex and age (thousands), annual [Data set]. Retrieved June 2025 from <https://rshiny.ilo.org/dataexplorer71/>

Version / Date Accessed: June 2025.

Filters Applied:

- **Dataset:** Labor force by sex and age (thousands) – Annual
- **Sex:** Total

- **Age group:** 15+ (Youth, adults)
- **Reference area:** All available countries and territories (229)
- **Region:** All regions (20)
- **Years:** full available range

Note on data versioning:

Lightcast provides access to its job-postings data through cloud-based services that are continuously updated. The database is dynamically maintained by Lightcast engineers, who apply weekly updates to incorporate new postings, remove duplicates and false positives, and refine data classifications.

According to Lightcast's official methodology [here](#):

- Job postings are revisited for up to 14 days after their initial collection to ensure completeness.
- A two-step deduplication process removes up to 80% of duplicate postings, using normalized title, company, and location fields over a 60-day window.
- Historical data are reclassified monthly, applying the most recent taxonomies and classification models to both current and past postings.
- Approximately 1% of outlier postings are filtered monthly, excluding misleading, discriminatory, or spam content.

As a result, dataset versions reflect the state of the Lightcast database as of their retrieval date and may differ slightly from more recent releases due to ongoing reclassification, deduplication, and data-quality improvements.

Citation

If you use these codes, please cite as:

Martins-Neto, Antonio; Liu, Yan; Khurana, Saloni; and Porras, Juan.

"Click, Code, Earn: The Returns to Digital Skills."

The World Bank, October 2, 2025.