Education Meets Opportunity Platform (EMOP)

Florida Department of Education

Comparing Florida Postsecondary Career Technical Education (CTE) Programs

Analysis Document

Authors:

Dr. A. Yonah Meiselman

Dr. Joe Long

October 20, 2023

RIPL | \ri'-pəl\ | **Research Improving People's Lives**
One Park Row, Suite 401, Providence, RI 02903
https://ripl.org

© 2023 Innovative Policy Lab, D.B.A. Research Improving People's Lives ("RIPL"). All Rights Reserved.

Purpose

This document serves as technical documentation of the Education Meets Opportunity Platform (EMOP) Return on Investment (ROI) dashboard postsecondary CTE program analysis to explain how the analysis dataset is constructed, how the analysis is run, and how the final calculations going into the results table are made. The intended audience is either an analyst seeking to replicate, edit, or utilize the analysis, or a researcher seeking to understand the exact methodology used in the analysis.

Outline

The description of the analysis is broken down into the following three sections:

Analysis Dataset Construction – The intermediate datasets[1] merged together to create one analysis dataset with student-by-program-completion as the level of observation. Many variables are combined or transformed to yield the outcomes of interest.

Sample Restrictions – The set of observations used in the analysis are selected in order to ensure that the outcomes and covariates (explained more in the regression adjustment section) are comparable across individuals. This includes, for example, keeping only the observations where the program was completed more than 8 quarters before the endpoint of the availability of earnings data.

Regression-Adjusted Average Outcomes Analysis – For each program (defined at several levels of granularity), the average of each outcome is calculated among all students completing that program, after adjusting for various controls. Exact details about the regression adjustment (including the definition of "regression-adjusted") are explained in this section.

Analysis Dataset Construction

This section covers part of the script "prog-ave-1-prep.r," which constructs the analysis dataset.

The analysis dataset, *prog_ave_prep.csv*, is created by merging the following intermediate datasets, which are the outputs from steps 5 and 6 in the data pipeline:

- *fcs_wages.csv*
- *wdis_wages.csv*
- *continuing_ed.csv*

The analysis dataset also incorporates the following external datasets (which would need to be updated for future analysis as these data change over time):

- A crosswalk between program codes and 10-digit CIP codes from both 2010 and 2020
- A list of CIP codes and their descriptions
- An inflation time series that shows the Consumer Price Index (CPI) covering the sample period
- A time series that shows the minimum wage in Florida during the sample period

The level of observation is student-by-program-completion. For each student, there is one observation for each program that student completed. Most students in the dataset complete only one program.

The analysis dataset contains a set of variables that can be divided into three sections:

- **Identifiers** – used to merge datasets together and to group individuals together into programs
  - *fl_anonymous_id, cip6, cip10, ed.level, completion_quarter*
- **Outcomes** – describe labor market activity and benefits receipt in the period after program completion

[1] The intermediate datasets are the outputs of the data pipeline (described in the "Runbook" section of the data pipeline document). They are the datasets created using the raw data provided which then go into the analysis.

    - *qemp_any, qemp_cum, wages_cum, wages_cond, qftmw_cum, qben_cum, ben_cum, cont.ed*
- **Covariates** – demographic indicators and pre-enrollment labor market history used as controls in calculating regression-adjusted averages
    - *male, high_school_education, first_time, race, mil, disability, age, earn.lag, qemp.lag*

A full data dictionary for the analysis dataset can be found in the Appendix.

Sample Restrictions

This section covers part of the script "prog-ave-1-prep.r," which constructs the analysis dataset.

1. **Panel length:** This analysis uses longitudinal data on individuals' earnings from before and after their program. The main restriction on the sample is that an observation can only be included if there are earnings data available for one year before program entry and two years after program completion. That is, observations are included if:

   (entry_quarter >= 2017-01-01) and (completion_quarter <= 2019-07-01)

   Note that observations are not excluded because they individually lack W-2 earnings. On the contrary, that is an important outcome. Rather, observations are excluded here only because they are from a time period in which earnings data are not available for anyone.

2. **Missing data:** Some observations lack information on a student's age. We exclude those observations.

Regression-Adjusted Average Outcomes Analysis

This section covers the scripts "prog-ave-2-reg.r" and "prog-ave-3-reorg.r," which run the regressions and format their results.

This section describes how we calculate the average outcomes for each program (or set of programs). These averages can be used to compare between postsecondary CTE programs. However, there is no comparison group of individuals who did not enter any such program. Therefore, the averages do not give us information about how postsecondary CTE programs affected outcomes relative to not entering into any postsecondary CTE program.

For each program, the raw average of each outcome and the regression-adjusted average of each outcome is calculated among all students completing that program. We define programs at five different granularity levels:

- 4-digit CIP code
- 4-digit CIP code x education level
- 6-digit CIP code
- 6-digit CIP code x education level
- 10-digit CIP code

At each of those five granularity levels, we create an indicator for each program defined at that level. Then we perform ordinary least squares (OLS) twice at each granularity level for each outcome. The outcomes include:

- *qemp_any, qemp_cum, wages_cum, wages_cond, qftmw_cum, qben_cum, ben_cum, cont.ed*

First, only the program indicators are included (no intercept). Second, both the program indicators and the covariates are included. The covariates include:

- Indicators for *male, high_school_education, first_time, race, mil, disability,* and *completion_quarter*
- Indicators for *age_bucket*, which is *age* divided into five buckets: 0-17, 18-25, 26-30, 31-65, 65+
- Interactions between *qemp.lag* and *earn.lag*

In econometric terms, we model outcomes as follows:

$$y_i = \sum_j x_{ij}\tilde{\beta}_j + w_i\gamma + \epsilon_i$$

Where y_i is an outcome for individual *i*, x_{ij} is a program indicator for program *j*, and w_i is an optional vector of covariates. We exclude the intercept, so there is no excluded program[2].

Let $\bar{w}$ be the average value of *w* over all individuals i. Our objects of interest are $\beta_j = \tilde{\beta}_j + \bar{w}\,\gamma$, the impact of each program plus the average covariate impact. This quantity corresponds to the outcome we would predict after program *j* is completed for a typical Florida student. For quick calculation of this quantity, we normalize all covariate values such that:

$$\widetilde{w}_i = w_i\gamma - \bar{w}$$

$$y_i = \sum_j x_{ij}\beta_j + \widetilde{w}_i\gamma + \epsilon_i$$

We estimate this last equation by OLS with and without standardized covariates $\widetilde{w}_i$. Then the estimated coefficients $\hat{\beta}_j$ correspond to the objects of interest.

The covariates allow us to isolate the impact of a program from the endogenous[3] correlation between which program is completed and the students' pre-program characteristics. Without the covariates, our estimates might be *biased*. For example, if a program happened to attract a set of students who already had high earnings before the program and if pre-program earnings are related to post-program earnings, then we might erroneously estimate that the "impact" of that program on earnings is high. With covariates, we remove that *bias*. When we refer to "regression-adjusted" averages, we are referring to the estimates from equations that include covariates.

Finally, we normalize the values of the estimated coefficients to be percentage differences from the mean. Specifically:

$$\mu_\beta = \frac{\sum_i \sum_j x_{ij}\hat{\beta}_j}{N}$$

$$\hat{\alpha}_j = \frac{\hat{\beta}_j - \mu_\beta}{\mu_\beta}$$

We report these percentage differences $\hat{\alpha}_j$.

Runbook

In order to run the analysis, there is a script called run-analysis.R which calls all three of the scripts listed above (*prog-ave-1-prep.r*, *prog-ave-2-reg.r, and "prog-ave-3-reorg.r).* It also lists all the input and output

[2] In a regression with categorical variables (such as programs), one can either include or exclude an intercept. Including an intercept requires that you choose a reference program against which the effects of the other programs are compared, and so all effects are relative outcomes. By excluding the intercept, what is calculated instead is the absolute outcome per program.
[3] Endogeneity refers to the relationship between an independent and dependent variable that is not the causal effect of the independent variable on the dependent variable.

file paths for ease of updating, which are also documented in the run-analysis.R tab of the Excel spreadsheet runbook-supp.xlsx, with the lines they occur and the names and descriptions of the files.

The analysis pipeline makes use of the following R packages:

- dplyr
- readr
- lubridate
- stringr

Appendix: Data dictionary for *prog_ave_prep.csv*:

- **Identifiers**
  - *fl_anonymous_id*
    - description: identifies a student
  - *cip6*
    - description: identifies the program completed at the 6-digit CIP code level
  - *cip10*
    - description: identifies the program completed at the 10-digit CIP code level
  - *ed.level*
    - description: identifies whether the program completed was a degree-granting program or a certificate-granting program
  - *completion_quarter*
    - description: identifies the quarter in which the program was completed
- **Outcomes**
  - *qemp_any*
    - description: 1 there is any quarter in which W-2 earnings in Florida were greater than zero, among the 8 quarters after the *completion_quarter*, 0 otherwise
  - *qemp_cum*
    - description: the number of quarters in which W-2 earnings in Florida were greater than zero, among the 8 quarters after the *completion_quarter*
  - *wages_cum*
    - description: the sum of all W-2 earnings in Florida over the 8 quarters after the *completion_quarter*, in 2012 dollars
  - *wages_cond*
    - description: if *qemp_any* = 1, *wages_cum*, 0 otherwise
  - *qftmw_cum*
    - description: the number of quarters in which W-2 earnings in Florida were greater than a “full-time” minimum wage (equal to the contemporary minimum wage times 520 hours per quarter) among the 8 quarters after the *completion_quarter*
  - *qben_cum*
    - description: the number of quarters in which the amount of SNAP and/or TANF benefits received was greater than zero, among the 8 quarters after the *completion_quarter*
  - *ben_cum*
    - description: the sum of all SNAP and/or TANF benefits received over the 8 quarters after the *completion_quarter*, in 2012 dollars
  - *cont.ed*
    - description: 1 if the individual enrolled in another program in the year following completion, 0 otherwise
- **Covariates**
  - *male*
    - description: 1 if the individual was male, 0 otherwise
  - *high_school_education*
    - description: 1 if the individual has a high school diploma or equivalent, 0 otherwise
  - *first_time*
    - description: 1 if the program completed was the student’s first time in post-secondary education
  - *race*
    - description: indicates with which racial or ethnic group the student identified
  - *mil*

        - description: indicates whether the individual was active duty military, veteran, or neither
    - *disability*
        - description: indicates whether the student had a disability
    - *age*
        - description: indicates the age of the student at program completion
    - *earn.lag*
        - description: the sum of all W-2 earnings in Florida over the 4 quarters before and including the *entry_quarter* (the quarter in which the student enrolled in the program they later completed), in 2012 dollars
    - *qemp.lag*
        - description: the number of quarters in which W-2 earnings in Florida were greater than zero, among the 4 quarters before and including the *entry_quarter* (the quarter in which the student enrolled in the program they later completed)