Merit Aid and STEM Project – Milestone #3

ECON 465 – M2

Burke, Patel, Singh, Smith, Yoon

31 October 2024

Reconciling Deficiencies from Milestone 2

The only discrepancies we received on Milestone 2 were in the classification of miscellaneous variables as numerical when they were actually numbers to represent categories (such as a code mapping numbers to birthplaces/states). These binary corrections were simply made by the instructor and do not merit reproduction here.

Describing Data Processing Plans

Several processing steps must occur after having downloaded the dataset above and before the team intends to conduct a full analysis and calculate regressions. For starters, all of the datasets were successfully merged upon download such that the raw data frame includes the full set of observations downloaded on one table, not multiple. As a preliminary processing step, the downloaded data sample was limited based on age and educational attainment within the range of consideration for the study. We eliminated study participants who did not attain a bachelor's degree to ensure relevance/analysis of only those affected by a merit-based college aid program.

Using the Table A1 supplement data frame, we also merged that with the main ACS dataset (on the variable degfieldd, the ACS code for each major) to include the BroadField category as a variable for each participant. This allows us to stratify based on STEM or non-STEM majors rather than only have distinct major names in the dataset per participant, by making another variable called STEM = 1 if STEM is the BroadField variable, 0 otherwise.

In general, we also included state-of-birth by year-of-birth time trends through the inclusion of an interaction term `i.state*birthyear`, which identified the birth state by birth year (`birthyear` was created by subtracting age from year of data collection) time trends. We also made dummy variables for state, and year of birth, individually. Across these three dummy categories, there were lots of dummies!

We create new categories to identify observations by age, educational attainment, and birthplace all together for the construction of our merit variables. These segmentations had to be conducted manually, as no one datapoint in ACS gives the date at which merit-based policies were instituted in each state. Thus, the years in which each merit program per state were enacted were individually researched via the sources indicated in the notes of Table 1 of the paper. We utilized the same stratifications of weak v.s. strong v.s. no merit (weak states are listed in the comments of Table 1, for example). We cross-checked the output of ChatGPT identifications of year in which merit programs were enacted with lists of such information in the paper's cited sources to ensure accuracy.

We identified the state and year of birth for each individual and made the following variables: 'strong' =1 if the person lived in a state where a strong merit program was enacted at some point, and 'weak' equal to 1 if the person lived in a state where a weak merit program was enacted at some point. We used existing variables such as weights for survey design and control variables like sex and race. They were made into dummy variables when necessary.

As a function of census year, individual age, and policy start year from our separately created dataset, we added three individual-level "merit access" variables to distinctly delineate who had access to merit-based financial aid in which year, in efforts to replicate the various columns of Table 3. Since the dataset is in units of person-year, each observation may have a different value for this dummy merit access variable, even if it is the same individual over the span of two years. The function for producing the variables ran generally as follows: if the policy start year was less than the individual's birth year (calculated like in the paper, year of survey minus age), then this individual had access to the given merit program.

These were correlated with the states' unique codes in the variable 'bpl', then used to create variables called 'strong_merit' and 'weak_merit' in the following fashion (just a few states shown for brevity):

```
strong_merit = (bpl==012 & birthyear+18>=1997) | (bpl==013 &
birthyear+18>=1993) | (bpl==021 & birthyear+18>=1999) | ...
```

Another variable titled `strong_or_weak_merit` was set equal to 1 if either 'strong_merit' or 'weak_merit' also equalled 1.

For each specific column of Table 3, we generated the "merit access" variable via the following mechanisms:

For column 1, individuals/observations received a 1 for `strong_merit` if their birthplace corresponded to a strong merit state and their age is below 18 at the time of policy implementation (as described above). Individuals received a 0 for treatment if they were in a non-merit state ('weak' & 'strong' == 0) or if they were in a strong merit state and had an age below 18 at time of policy ('strong_merit'==0). All others would not be included in analysis.

For column 2, the same classifications as column 1 were utilized, but this time, individuals were included in analysis if 'weak' == 1 (anyone in a weak merit state).

For column 3, we needed to utilize the large list (hundreds) of dummy variables for state-year interactions, where each state-year pairing was a new dummy variable for which we had to control.

For column 4, we only identified observations as 1 for treatment if they were in a strong merit aid state and turned 18 after the program was implemented, and as 0 only if they are also in a strong merit state but too young (so 'weak' ==1 were not included at all, as well as anyone who did not have 'strong'==1).

For column 5, all states were included in the sample (no limitations based on strength of merit aid program) and students who received a 1 for their `weak_merit` variable was put into the treatment group. In other words, the variable `strong_or_weak_merit` became the treatment variable.

Initial Analysis

To replicate the key regression from the assigned paper, we will conduct an econometric analysis using the following Linear Probability Model:

$$P(Y_{isct}=1)=\beta_1 sex+\beta_2 race+\beta_3 hispa+\beta_4 age\ +\ \beta_5 educ+\beta_6 Merit_{sc}+\delta_s+\gamma_c \\ +\pi_s T_{sc}+\epsilon_{isct}$$

Where $P(Y_{isct}=1)$ is the probability that individual i , from state s , cohort c , and survey year t , completed a STEM degree. The variables `sex, race, hispan, age`, and `educ` includes control variables such as sex, race, age, Hispanic origin, and educational background. The model includes state fixed effects δ_s to control differences across states and birth cohort fixed effects γ_c to account for birth-year cohort variations. The $\pi_s T_{sc}$ term accounts for state-of-birth by year-of-birth time trends, basically `i.state*year`. The error term ϵ_{isct} captures unobserved factors. The unit of observation is person-year.

To account for the different specifications in Table 3, we will create separate merit variables. The variable $Merit_{sc}$ will change based on which version of the variable we are using. First, we will create a variable $Merit_{strong}$, where $Merit_{sc}$ will only include only individuals exposed to strong merit aid programs assigning a value of 1 if the individual was born in a state with a strong program and turn 18 after its implementation and 0 otherwise. Second, we will create $Merit_{both}$, where $Merit_{sc}$ will include both strong and weak merit aid programs assigning value of 1 for exposure to either type or 0 otherwise. Third we will create a regional comparison version, $Merit_{region}$, where $Merit_{sc}$ will compare individuals from merit aid states to those from non-merit aid states within the same region. Lastly, we will create $Merit_{prepost}$, where $Merit_{sc}$ will focus on individuals from strong merit aid states, assigning a value of 1 if they turned 18 after the program was introduced, and 0 if they turned 18 before. These versions of $Merit_{sc}$ will be used in different regressions. This is a reiteration of the data processing tasks, more specifically identifying how we will utilize it in our analysis regressions.

The analysis will be estimated using Linear Probability Models, not probit or logit as the authors state they do not use these as the coefficients on LPM are easier to interpret. We know that this will answer the primary research question of our paper, how merit-based aid program access affects acquisition of a STEM-based major, as the coefficient on the Merit variable term will be able to identify the effect (percentage change) of STEM majors given the successful implementation of the unbiased LPM. The standard errors will be clustered by state of birth to account for within-state similarities. We will not utilize Conley-Taber confidence interval procedure as this was specified in the assignment guidance. The state-of-birth and year-of-birth fixed effects variables can allow the model to be considered a difference-in-differences estimation.

Finally, we used Stata for all of our data analysis regressions, of which the only thing changing between regressions for each column should be the `Merit` variable, which includes/excludes certain groups of states' programs' strength. All of the coefficients necessary to reproduce the graph should be evident in the model output on Stata.

To assess the impact of state merit aid programs on students' selection of a STEM major, we ran a series of regressions across five models, each with slightly different criteria for the treatment and control groups. In column 1, the sample is limited to individuals born in strong merit states without including weak merit states, and the `strong_merit` coefficient shows a decrease of 1.32 percentage points in the likelihood of selecting a STEM major. In column 2, weak merit states are included as part of the control group, showing a slightly smaller but still negative effect of -0.73 percentage points. Column 3 includes state-year dummy variables to account for state-specific trends over time, producing an estimated effect of -1.23 percentage points. Column 4, which includes only strong merit states, reflects a similar decrease of around 1.29 percentage points, while column 5 includes both strong and weak merit states in the treatment group, showing a reduced effect of -0.38 percentage points. This consistent negative association between `strong_merit` and STEM choice across all models suggests that students in states with strong merit programs are less likely to select STEM majors, aligning with the research hypothesis regarding the possible influence of GPA requirements on major selection.

Interpretation and Comparison

Our results are very similar to those of Sjoquist and Winters. In their Table 3, the authors show reductions in the likelihood of choosing a STEM major for students in different types of merit aid program groups. In column 1, which includes only people born in strong merit states exposed to strong programs (excluding weak states), they find a decrease of about 1.3 percentage points. Column 2, where weak merit states are added to the control group, also shows a 1.3 percentage point drop. In column 3, where state-year interactions are included to account for trends over time, the effect is a slightly smaller 1.2 percentage points. Column 4, focusing only on students from strong merit states, shows a decrease of about 1.4 percentage points. Column 5, which treats both strong and weak merit states as the treatment group, shows a smaller decrease of 0.5 percentage points.

These match the decreases we observed in our own analysis, where several models show a similar trend of fewer students choosing STEM majors in response to merit aid programs. Each of our models is highly statistically significant, with p-values close to 0.000. Sjoquist and Winters also report statistically significant results, which adds further support to the idea that strong merit aid programs reduce STEM participation. This similarity in both the size of the effect and the statistical certainty across both studies suggests that strong merit programs do indeed reduce the likelihood of students selecting a STEM major.

Some minor differences in the exact values of our coefficients compared to the original study may arise due to sample variations. This may include differences in demographics or geographic regions. Additionally, slight variations in variable definitions (how race categories or program strengths are classified) and differences in statistical adjustments such as state and year controls may also impact specific estimates. However, these factors highlight that while exact figures can vary slightly, the overall finding remains consistent. Strong merit aid programs are associated with a lower likelihood of students choosing STEM majors.

Subsequent Analysis

Sjoquist and Winters conducted additional analysis to examine shifts in major selection among students who were less likely to choose STEM due to merit aid programs. Specifically, they explored whether students in states with strong merit aid programs shifted toward non-STEM fields, such as business, social sciences, or liberal arts, which may carry fewer academic challenges compared to STEM. Their findings indicate that students exposed to merit aid were indeed more likely to select these fields. This suggests that students may consider the ability of maintaining GPA requirements when choosing their major.

This additional analysis highlights the broader impact of merit aid programs on students' academic choices. While these programs increase college affordability, they may also make students gravitate toward less demanding fields. This will help them potentially meet GPA requirements for scholarship retention. By examining students' shifts away from STEM, Sjoquist and Winters provide a perspective on how merit aid influences educational paths. They show that students are still likely to complete degrees but may prioritize maintaining scholarship eligibility over pursuing challenging fields like STEM.

The further analyses conducted in the tables we did not study include whether or not the implementation of the merit aid programs affected other degree fields outside of STEM. They essentially conducted the regressions performed on the STEM model to various other major categories in the humanities. The concern that motivated such analysis was that they wanted to gauge the impact of the STEM increase/decrease caused by the merit aid programs in various other fields, so essentially answering the question of where do the STEM dropouts go/end up majoring in. The differences between Beta-1 coefficients for these various humanities regressions can answer the relative strength of change or popularity gain of that given major category, answering the concern arisen.

Appendix

Column 1 code:

```
reg STEM strong_merit sex race age hispan if weak==0
```

Column 2 code

```
reg STEM strong_merit sex race age hispan
```

Column 3 code

```
gen state_year_dummy = bpl*birthyear

reg STEM strong_merit sex race age hispan state_year_dummy if weak==0
```

Column 4 (Dummies included as our output here was closer to paper results)

```
reg STEM strong_merit sex race age hispan dummy_state_* dummy_birthyear_* if strong==1
```

Column 5

```
reg STEM strong_or_weak_merit sex race age hispan dummy_state_* dummy_birthyear_*
```

Documentation Statement: Some documentation used for this milestone included the IPUMS dataset explainer tool for our downloaded data, the project paper itself, its references, and our Milestone 1 & 2 commentary/feedback. We also utilized ChatGPT for the identification of merit years, the transcript of which conversation can be read here.