Methodological Appendix

Are Community-College Students Increasingly Choosing High-Paying Fields of Study? Evidence from Massachusetts

I. Introduction

The field of education is replete with questions about the occurrence and timing of students' receipt of important academic credentials such as a high-school diploma or an associate or bachelor's degree. Addressing such questions is challenging because of the ubiquitous presence of censoring. For instance, in any sample of students who are followed over time from their entry into a community college until they earn an associate degree (AD), say, there will almost certainly be students present who do not earn an associate degree while they are being observed. The longitudinal records of such students are said to be *right-censored*. If the period of observation is short, or if times-to-AD are long, then cases with censored records may even predominate in the sample.

Unfortunately, unbiased estimates of overall population times-to-event cannot be obtained by analyzing observed times-to-event in the subsample of uncensored cases – that is, by averaging among the students who actually earned an AD while being observed. This is because some, or even many, of the students with censored records may eventually go on to attain an AD after the period of observation has ended. These censored students will then possess times-to-event that are equally legitimate yet longer than their colleagues with uncensored records. Consequently, estimating average time-to-AD among only uncensored students in the sample will almost certainly underestimate the average time-to-AD in the entire population (i.e., the estimate of time to AD will be negatively biased).

There are several ways to deal with this problem. The most obvious -- and seemingly simple solution -- is to modify the research design and follow every individual in the sample until they have actually experienced the event of interest and earned the AD or died. Unfortunately, this is also the most untenable design. Some unknown yet perhaps substantial part of the entrant-sample may simply labor on, year after year, perhaps taking ten, twenty, thirty years to earn their AD, providing institutional rules permit. Everyone in the sample would then need to be followed until such time as they had all earned the AD or died, contributing ever-more extreme datapoints to the average time-to-event or to the regression coefficient that represents it in more complex analyses. Each succeeding datapoint whose value was added to the average lifetime then makes an even-more-extreme contribution to the final estimated average lifetime. And it is only if the observer outlived all of the sampled subjects and stayed on the job for that duration that the sample-average lifetime could be known without bias.

One sensible alternative to such a potentially interminable task -- and one that we have adopted here -- is to reframe the research question so that its "answer" can be proffered in some alternative sensible and readily interpretable metric, such as the metric of probability, whose values – in this context -- can be estimated unbiasedly even in the presence of censoring. For instance, instead of waiting potentially interminably for all sampled students to earn an AD in order to estimate a legitimate unbiased average time-to-AD, one can simply dichotomize the continuous times-to-AD obtained in the sample at some reasonable criterion duration -- such as four, five, or six years, say – and then describe patterns of AD-attainment among students by estimating the probability that they had attained an AD by the end of that agreed-upon time window. In our study, for instance, we chose to dichotomize AD-attainment window at six years (with a value of 1 if a student had attained the AD within a six-year window from MACC entry,

and 0 if they had not), and so our research question became a probabilistic contrast between students who did, versus those who did not, earn an AD within the target six years. In our case, then, the average value of our newly dichotomized (0/1) outcome provides an unbiased estimate of the probability that students attained the AD within a six-year window following their entry into a community college.

However, given that our fundamental intent was to investigate *cross-cohort trends* in the outcome and that our observational data record terminates at the end of the Spring term of the 2021-22 school year, there is an obvious tradeoff that we were forced to accept in choosing the temporal cutoff for dichotomization. That is, the longer the time-period within which we chose to observe students working toward an AD, the fewer full cohorts of students we could define and include in the analysis. For example, if we had permitted students an eight-year period from MACC-entry in which to earn an AD, we could then only include students from the 2005 through 2014 entry cohorts in our analytic sample before the full window of AD-opportunity becomes truncated in the later cohorts. In our case, preliminary survival analyses of student records from early cohorts in our data repository who actually earned an AD in either Healthcare or in STEM, relatively few students took longer than six years to do so. Consequently, we adopted six years as our temporal cutoff for dichotomization, and focused on trends across the 2005 through 2016 entry cohorts in the probability of AD-attainment within six years of entry for these entry cohorts. In the main body of the paper, we describe this dataset and present our findings. In what follows here, we define the variables featured in our analyses and describe the method of multinomial logistic regression analysis used to address our research question.

II. Measures

After dichotomization of the academic career at 6 years, every student in each entry

cohort contributed a single observation (row) to the analytic dataset, which indicated whether they had attained the associate degree within six years after entry into a MACC and listed the values of selected student attributes that we hypothesized would predict the outcome. Below, we define these variables.

II.1: Outcome Variable

In our multinomial logit analyses, our outcome was the trichotomous variable *AD*, which took on categorical values of either 0, 1, or 2. *AD* was assigned a value of 1 if the student had earned an associate degree in a healthcare program within six years of entry. It took on a value of 2 if the student had received an associate degree in a STEM program within the same period. If the value of outcome *AD* was zero (0), then the student had not received an associate degree in either field within six years of MACC entry.

II.2: Question Predictors

In our first research question, we sought to contrast cross-cohort trends in the probability that students earned an associate degree in either a healthcare or STEM field within six years after MACC entry, as a function of important student attributes. Thus, we defined the following predictors to represent entry cohort, MACC and hypothesized student attributes in our multinomial logit models:

- ***COHORT*** is a vector of dichotomous (0/1) predictors, named *C5* through *C16*, that distinguish the cohorts of students who entered a MACC annually in the fall of years 2005 through 2016 respectively, with the value of each component dichotomy in the vector set to a value of 1 in the corresponding entry year, and zero (0) otherwise. In specifying our multinomial logit models, we elected to omit element *C10* from the ***COHORT*** vector as a predictor, thereby forcing the 2010 entry-year to become the origin of the cohort effects

being modeled. Thus, cohort predictor-vector {*C5*, *C6*, *C7*, *C8*, *C9*, *C11*, *C12*, *C13*, *C14*, *C15*, *C16*} is present in each of our hypothesized statistical models to provide a general specification of the main ("fixed") effects of entry cohort on the receipt of an associate degree.

- ***PCOHORT***: For purposes of parsimony in specifying subsequent statistical interactions between substantive student predictors (defined below) and *COHORT*, we also defined as an additional continuous predictor a centered linear function of cohort (*LINCOH*).

In addressing our principal research questions, we sought to investigate how cross-cohort differences in the six-year probability of earning an associate degree in healthcare or STEM differed systematically by selected student attributes, including student gender, student race/ethnicity, family income, and student mathematics achievement in high-school, in that order of precedence. To test these hypotheses, we created the following question predictors:

- *FEM* is a time-invariant dichotomous predictor, with a value of one (1) if the student had declared themself female and zero (0) if they indicated they were male.
- ***RACE*** is a vector of five time-invariant dichotomous predictors, {*ASIAN*, *BLACK*, *HISPANIC*, *WHITE*, *OTHER*}, each coded one (1) if a student declared they were of the corresponding race/ethnicity. In specifying our statistical models, we omitted predictor *WHITE* from this vector to create a reference category. In interpreting differences among the probabilities of degree-attainment by levels of these predictors, keep in mind that there is substantial variation in heritage and resources among students *within* each broad racial/ethnic group enrolled in Massachusetts' community colleges. For example, as one community-college president reminded us, the students who were of Asian descent and

enrolled in some Massachusetts' community colleges came predominantly from families that had arrived in the USA recently, from Cambodia and Laos. These families tended to have fewer resources and weaker English-language skills than did the families of Asian descent who had emigrated from China and India at least a decade earlier and were now enrolled disproportionately in other Massachusetts' community colleges. Similar granularity also existed within the broader "Hispanic," "Black" and "White" ethnicities. Unfortunately, we did not possess the information needed to distinguish these finer subtleties of student race/ethnicity in our statistical models.

- *LOWINC* is an indicator of the income of the family in which the student was raised while in high school. It is a dichotomous predictor that takes on a value of 1 if the student was eligible for a free- or reduced-price lunch in 10^{th} Grade, zero otherwise.
- *TOPHALF* is a dichotomous measure of student mathematical skills, derived from the score that the student achieved on the MCAS mathematics achievement test, which is taken by virtually all students in Massachusetts public high schools in 10^{th} Grade. This score was published by the State in three two versions (raw and scaled). However, because of problems with scale drift in the scaled scores, we followed the Massachusetts Department of Education's recommendation (2023) to model the relationship with associate-degree receipt and student mathematics achievement using the standardized measure of student achievement. Standardized scores possess a mean of zero and a standard deviation of unity in every testing year, in the population of all first-time 10^{th} Grade test-takers in Massachusetts. Thus, as a result, any true annual gain in average student mathematics from cohort to cohort has been removed from the scores by standardization. Effectively, by standardization, student mathematics achievement has been transformed into an indicator

of the students' *relative placement* (rank order) within the mathematics skill distribution of all tenth-grade students in their own test-taking cohort. Given that student mathematics achievement scores were therefore no longer equatable longitudinally from cohort to cohort, we created a dichotomous predictor *TOPHALF* to describe whether each student's score fell into either the top (=1) or bottom (=0) halves of the 10^{th} Grade MCAS mathematics score distribution for their own test-taking cohort. As will be noted below, this has implications for subsequent interpretation of the impact of mathematics achievement on associate-degree receipt.

II.3: Covariates

Finally, each student in our analytic sample began their post-secondary education at one of fifteen community colleges in Massachusetts. We identified their initial community college selection by defining the following covariate:

- ***MACC*** is a vector of dichotomous (0/1) predictors, named *M1* through *M15*, to distinguish among Massachusetts community colleges. In our statistical models, we elected to drop element *M3* from this vector, thereby implicitly choosing Bunker Hill Community College, one of the largest community colleges in the Massachusetts' system, as a reference category.

III: Data Analyses

To address our first research question, we fitted a sequential taxonomy of multinomial logistic-regression models in which polychotomous outcome *AD* was regressed sequentially on the cohort predictor-vector and selected question predictors, and on any statistical interactions among them that proved to be required, controlling for the fixed effects of community college. In *Equation A1*, we specify the initial "baseline" multinomial logit-regression model with which we

began the hypothesized taxonomy of models. Notice that it contains only the fixed effects of both entry cohort and MACC. In the model, $p_j^{(r)}$ represents the probability of associate-degree receipt in a particular field *r* (*r* = 0, 1, 2) within six years of MACC entry, for student *j* (= 1, …, *N*). In specifying this baseline model, we have followed the standard practice of specifying the outcome as logit functions – that is, the natural logarithm of a ratio of population probabilities of AD attainment – of the hypothesized predictors (Hosmer & Lemeshow, 2000). The model is indexed to subsume the two sub-models needed to span the Healthcare and STEM target-degree space (Healthcare: r = 1; STEM: r = 2). Finally, note that the "no-AD" condition (r = 0) provides the denominator of the logit-probability ratio that serves as the dependent variable in the model specification:

$$ln\left(\frac{p_j^{(r)}}{p_j^{(0)}}\right) = \alpha^{(r)} + \beta_1^{(r)}\boldsymbol{COHORT}_j^{(r)} + \beta_2^{(r)}\boldsymbol{MACC}_j^{(r)} \qquad \text{(A1)}$$

Notice that this hypothesized baseline model contains several classes of parameters. First, intercept vector $\boldsymbol{\alpha}^{(r)}$ provides a set of intercepts to anchor the cross-cohort trajectories of students in each of the fields of study. Second, slope-parameter vectors, $\boldsymbol{\beta}_1^{(r)}$ and $\boldsymbol{\beta}_2^{(r)}$ represent the fixed effects of entry cohort and community college. As noted above, these latter effects were retained throughout all subsequent models fitted in our analyses.

Then, to address our first research question specifically, we added selected predictors systematically to the baseline model, beginning with those predictors that described basic student demographic characteristics: (a) gender (*FEM*), and (b) race/ethnicity (***RACE***), as follows:

$$ln\left(\frac{p_j^{(r)}}{p_j^{(0)}}\right) = \alpha^{(r)} + \beta_1^{(r)}\boldsymbol{COHORT}_j^{(r)} + \beta_2^{(r)}\boldsymbol{MACC}_j^{(r)} + \beta_3^{(r)}\boldsymbol{FEM}_j^{(r)} + \beta_4^{(r)}\boldsymbol{RACE}_j^{(r)} \qquad \text{(A2)}$$

In fitting this model, we also tested for the presence of statistically significant interactions between the two new demographic predictors themselves and among each of them and the polynomial representations of entry-level cohort, *PCOHORT*. We retained all interactions that were statistically significant in subsequent models.

Finally, in the third and fourth rounds of model-building, we added predictor *INCOME* which described the student's family income in high-school, and also our dichotomous indicator of student mathematics achievement, *TOPHALF*, as the final predictors of the probability of AD-attainment, as follows:

$$ln\left(\frac{p_j^{(r)}}{p_j^{(0)}}\right) = \alpha^{(r)} + \beta_1^{(r)}\boldsymbol{COHORT}_j^{(r)} + \beta_2^{(r)}\boldsymbol{MACC}_j^{(r)} + \beta_3^{(r)}\boldsymbol{FEM}_j^{(r)} + \beta_4^{(r)}\boldsymbol{RACE}_j^{(r)} + \beta_5^{(r)}\boldsymbol{INCOME}_j^{(r)} + \beta_6^{(r)}\boldsymbol{TOPHALF}_j^{(r)} \quad \text{(A3)}$$

Again, we tested and retained any statistically significant interactions between the added predictors and predictors already present in the model and between the added question predictors themselves. We list parameter estimates, along with corresponding precision and fit statistics, for our final fitted model in *Table A1*.

Table A1: Final fitted multinomial logistic-regression model summarizing the relationship between receipt of an associate's degree in either the Healthcare or STEM fields within six years of entry into a Massachusetts Community College and selected student attributes, including gender, race/ethnicity, family income and mathematics achievement (N = 81,911, *Wald* χ^2 = 3059, df = 70, p = .000)

Predictors	***Field of Associate-Degree Study***					
	HEALTHCARE			***STEM***		
	Parameter Estimate	***Std. Error***	***p-value***	***Parameter Estimate***	***Std. Error***	***p-value***
Intercept	-6.509***	0.207	.000	-3.889***	0.132	.000
Entry Cohort:						
COHORT	*<Included>*		.001	*<Included>*		.000
Student Gender:						
FEM	2.234***	0.102	.000	-1.532***	0.091	.000
FEM×LINCOH				0.035**	0.017	.039
Student Race/Ethnicity:						
ASIAN	0.969***	0.247	.000	0.611***	0.095	.000
BLACK	-0.379***	0.139	.006	-0.643***	0.120	.000
HISPANIC	-0.281***	0.095	.003	-0.125	0.074	.089
OTHER	-0.907***	0.338	.007	-0.119	0.200	.552
ASIAN*×FEM*	-0.952***	0.298	.001			
BLACK×FEM				0.599***	0.220	.006
Family Income						
LOWINC	0.239	0.162	.141	-0.126**	0.058	.030
LOWINC×LINCOH				-0.032**	0.015	.034
*LOWINC×**FEM***	-0.495***	0.173	.004			
Student Math Achievement:						
TOPHALF	0.604***	0.057	.000	0.898***	0.051	.000
TOPHALF×FEM				0.233*	0.116	.045
Institution:						
MACC	*<Included>*		.000	*<Included>*		.000

*p<.05, ** p<.01,*** p<.001

References

Hosmer, D.W. & Lemeshow, S. (2000). Applied Logistic Regression (2nd Edition). NY: John Wiley & Son.