Analysis File

Shanshan Liu

**Introduction**: The primary purpose of this data exploration project is to investigate whether the release of earnings in the College Scorecard dataset in early September, 2015 has led to a shift in the preferences of prospective students towards universities with high-income graduates. In the research, the Google trend index has been used as a proxy of people’s preference towards colleges. The Google Trends data provides insights into the relative search popularity of keywords over time and the data can be used to discern patterns in search activity associated with universities. Through analysis and interpretation of these data compared with search popularity (search frequency), it can be analyzed whether the release of earnings in the Scorecard dataset has resulted in students’ interest turning towards high-income universities. This project has focused on the universities which primarily grant bachelor’s degrees.

**Methodology**: The following DiD regression is analyzed using Ordinary Least Square.

*Indexit* =*β*0​+*β*1 ​*HighIncomei+β*2 ​*Sept2015it+β*3 ​*HighIncomei\*Sept2015it* + *β* Ci

+ *αs + αt* + *ε*

Where subscript *i* stands for college, subscript *t* stands for month, and *s* stands for state. *Indexit* is the standardized Google Trend Index aggregated for each college and in each month. *HighIncomei = [high75i; highmeani]* is a set of dummy variables that represent colleges with high-income students. *High75i equals one if* the median earnings of graduates from college *i* 10 years after graduation ishigher than the 75th percentile in the sample. *Highmeani* equals one if the median earning of graduates from college *i* 10 years after graduation is higher than the average income. Both dummies are examined in this project. *Sept2015it* is a dummy variable that equals one if the index was recorded in or after September 2015. The coefficient on the interaction term, *β*3, is the coefficient of interest, which tells the impact of releasing earning data on online-searching activity of college-related keywords.

*αs* and *αt*are state and year fixed effects respectively. Different states may have different economic environments and social policies, which could influence university preferences. Adding state fixed effects in the regression model helps control for these state-level characteristics, allowing for a more accurate assessment of the impact of earnings releases on university preferences. As time progresses, changes in economic environments and social policies may occur, which could also affect university preferences and search activities. Introducing year fixed effects into the regression model helps control for these time trends.

*Ci* stands for control variables. I have controlled the types of ownership of schools and included dummies for public schools and private not-for-profit. The third type of ownership, private for-profit, is omitted to avoid collinearity. I believe that public schools and private schools may simultaneously influence graduate income and people’s interest in searching for related universities. Controlling for them can help eliminate the influence of these factors on preferences for high-income universities, making the impact of university scorecards on university preferences more accurate.

## Libraries

In the following code chunk, load all the libraries you will need:

#load packages  
library(rio)

Warning: package 'rio' was built under R version 4.2.3

library(stringr)  
library(lubridate)

Attaching package: 'lubridate'

The following objects are masked from 'package:base':  
  
 date, intersect, setdiff, union

library(dplyr)

Warning: package 'dplyr' was built under R version 4.2.3

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(fixest)

Warning: package 'fixest' was built under R version 4.2.3

## Load Data

# Specify the directory containing my files  
setwd("C:/5300 Applied Econometrics/Data Exploration Assignment/")  
selected\_df <- rio::import('selected\_df.csv')

## Run regression model

# regression model with fixed effect  
#model2 <-feols(index\_sdd\_shmonth ~ sept2015 + high75 + sept2015:high75 + public + private\_np , data = selected\_df, fixef = 'year',vcov = 'hetero')  
model1 <-feols(index\_sdd\_shmonth ~ sept2015 + high75 + sept2015:high75 + public + private\_np + factor(year) + factor(STABBR) , data = selected\_df, vcov = 'hetero')  
summary(model1)

OLS estimation, Dep. Var.: index\_sdd\_shmonth  
Observations: 30,285   
Standard-errors: Heteroskedasticity-robust   
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.94911696 1.144632 2.576476 9.9860e-03 \*\*   
sept2015 -0.00000323 0.206134 -0.000016 9.9999e-01   
high75 0.36827697 0.148915 2.473062 1.3402e-02 \*   
public 0.00862717 0.180312 0.047846 9.6184e-01   
private\_np 0.02163618 0.156199 0.138517 8.8983e-01   
factor(year)2014 -2.56388342 0.132108 -19.407507 < 2.2e-16 \*\*\*  
factor(year)2015 -5.31264637 0.138329 -38.405757 < 2.2e-16 \*\*\*  
factor(year)2016 -4.97271866 0.260088 -19.119388 < 2.2e-16 \*\*\*  
factor(STABBR)AL 0.00630622 1.179522 0.005346 9.9573e-01   
factor(STABBR)AR 0.01335104 1.203516 0.011093 9.9115e-01   
factor(STABBR)AZ 0.00951905 1.295636 0.007347 9.9414e-01   
factor(STABBR)CA 0.01233726 1.154711 0.010684 9.9148e-01   
factor(STABBR)CO 0.01366456 1.185512 0.011526 9.9080e-01   
factor(STABBR)CT 0.02019051 1.186376 0.017019 9.8642e-01   
factor(STABBR)DC 0.00234991 1.186502 0.001981 9.9842e-01   
factor(STABBR)DE 0.00203400 1.464353 0.001389 9.9889e-01   
factor(STABBR)FL 0.03625421 1.158667 0.031290 9.7504e-01   
factor(STABBR)GA 0.02641656 1.165855 0.022659 9.8192e-01   
factor(STABBR)HI -0.00029574 1.460716 -0.000202 9.9984e-01   
factor(STABBR)IA 0.02185131 1.166816 0.018727 9.8506e-01   
factor(STABBR)ID 0.02251282 1.544901 0.014572 9.8837e-01   
factor(STABBR)IL 0.03720290 1.159200 0.032094 9.7440e-01   
factor(STABBR)IN 0.01856972 1.181205 0.015721 9.8746e-01   
factor(STABBR)KS 0.04922381 1.193441 0.041245 9.6710e-01   
factor(STABBR)KY 0.02071410 1.192247 0.017374 9.8614e-01   
factor(STABBR)LA 0.03569788 1.162571 0.030706 9.7550e-01   
factor(STABBR)MA 0.02958443 1.153147 0.025655 9.7953e-01   
factor(STABBR)MD 0.02630824 1.183235 0.022234 9.8226e-01   
factor(STABBR)ME 0.01686570 1.212313 0.013912 9.8890e-01   
factor(STABBR)MI 0.02807204 1.185637 0.023677 9.8111e-01   
factor(STABBR)MN 0.01595869 1.189790 0.013413 9.8930e-01   
factor(STABBR)MO 0.02374555 1.159902 0.020472 9.8367e-01   
factor(STABBR)MS 0.03297414 1.218928 0.027052 9.7842e-01   
factor(STABBR)MT 0.02251282 1.400987 0.016069 9.8718e-01   
factor(STABBR)NC 0.03137463 1.166742 0.026891 9.7855e-01   
factor(STABBR)ND 0.04346464 1.223149 0.035535 9.7165e-01   
factor(STABBR)NE 0.01523158 1.185813 0.012845 9.8975e-01   
factor(STABBR)NH -0.01184414 1.435127 -0.008253 9.9342e-01   
factor(STABBR)NJ 0.02396798 1.195020 0.020057 9.8400e-01   
factor(STABBR)NM 0.04051704 1.581779 0.025615 9.7956e-01   
factor(STABBR)NV 0.00397235 1.169575 0.003396 9.9729e-01   
factor(STABBR)NY 0.01898620 1.137743 0.016688 9.8669e-01   
factor(STABBR)OH 0.02048243 1.156024 0.017718 9.8586e-01   
factor(STABBR)OK 0.02536994 1.216816 0.020849 9.8337e-01   
factor(STABBR)OR 0.02612132 1.178294 0.022169 9.8231e-01   
factor(STABBR)PA 0.03569375 1.144121 0.031198 9.7511e-01   
factor(STABBR)PR 0.00458186 1.165249 0.003932 9.9686e-01   
factor(STABBR)SC 0.02171921 1.177930 0.018438 9.8529e-01   
factor(STABBR)SD 0.01596699 1.260801 0.012664 9.8990e-01   
factor(STABBR)TN 0.03128712 1.157766 0.027024 9.7844e-01   
factor(STABBR)TX 0.00924591 1.151388 0.008030 9.9359e-01   
factor(STABBR)UT 0.00110571 1.274455 0.000868 9.9931e-01   
factor(STABBR)VA 0.02450431 1.158153 0.021158 9.8312e-01   
factor(STABBR)VT 0.03711112 1.189429 0.031201 9.7511e-01   
factor(STABBR)WA -0.00274375 1.249586 -0.002196 9.9825e-01   
factor(STABBR)WI 0.01158097 1.210625 0.009566 9.9237e-01   
factor(STABBR)WV 0.02969483 1.277124 0.023251 9.8145e-01   
sept2015:high75 -2.29259145 0.329501 -6.957777 3.5272e-12 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
RMSE: 8.29483 Adj. R2: 0.066133

# regression model with interaction highmean   
#model5 <- feols(index\_sdd\_shmonth ~ sept2015 + highmean + sept2015:highmean + public + private\_np, fixef = 'year', data = selected\_df, vcov = 'hetero')  
model2 <- feols(index\_sdd\_shmonth ~ sept2015 + highmean + sept2015:highmean + public + private\_np + factor(year) + factor(STABBR) , data = selected\_df, vcov = 'hetero')  
summary(model2)

OLS estimation, Dep. Var.: index\_sdd\_shmonth  
Observations: 30,285   
Standard-errors: Heteroskedasticity-robust   
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.949533 1.144265 2.577665 9.9517e-03 \*\*   
sept2015 0.001330 0.213996 0.006215 9.9504e-01   
highmean 0.210103 0.122581 1.713989 8.6541e-02 .   
public 0.008155 0.176866 0.046110 9.6322e-01   
private\_np 0.021098 0.151176 0.139557 8.8901e-01   
factor(year)2014 -2.564103 0.132124 -19.406759 < 2.2e-16 \*\*\*  
factor(year)2015 -5.312866 0.138337 -38.405343 < 2.2e-16 \*\*\*  
factor(year)2016 -4.972938 0.260611 -19.081874 < 2.2e-16 \*\*\*  
factor(STABBR)AL 0.006937 1.179712 0.005880 9.9531e-01   
factor(STABBR)AR 0.013753 1.203574 0.011427 9.9088e-01   
factor(STABBR)AZ 0.011661 1.294611 0.009007 9.9281e-01   
factor(STABBR)CA 0.014538 1.155633 0.012580 9.8996e-01   
factor(STABBR)CO 0.015219 1.185878 0.012834 9.8976e-01   
factor(STABBR)CT 0.021174 1.186350 0.017848 9.8576e-01   
factor(STABBR)DC 0.004117 1.185531 0.003473 9.9723e-01   
factor(STABBR)DE 0.004242 1.464522 0.002897 9.9769e-01   
factor(STABBR)FL 0.037386 1.158473 0.032272 9.7426e-01   
factor(STABBR)GA 0.027332 1.166127 0.023438 9.8130e-01   
factor(STABBR)HI 0.000350 1.466694 0.000239 9.9981e-01   
factor(STABBR)IA 0.022840 1.166961 0.019572 9.8438e-01   
factor(STABBR)ID 0.022535 1.544898 0.014587 9.8836e-01   
factor(STABBR)IL 0.038942 1.159408 0.033588 9.7321e-01   
factor(STABBR)IN 0.020013 1.181092 0.016944 9.8648e-01   
factor(STABBR)KS 0.049263 1.193443 0.041278 9.6707e-01   
factor(STABBR)KY 0.021398 1.192609 0.017942 9.8569e-01   
factor(STABBR)LA 0.036779 1.163056 0.031622 9.7477e-01   
factor(STABBR)MA 0.031840 1.152557 0.027625 9.7796e-01   
factor(STABBR)MD 0.029528 1.186067 0.024896 9.8014e-01   
factor(STABBR)ME 0.016961 1.210433 0.014013 9.8882e-01   
factor(STABBR)MI 0.029029 1.185000 0.024497 9.8046e-01   
factor(STABBR)MN 0.018610 1.191639 0.015617 9.8754e-01   
factor(STABBR)MO 0.024109 1.159839 0.020786 9.8342e-01   
factor(STABBR)MS 0.033206 1.219013 0.027240 9.7827e-01   
factor(STABBR)MT 0.023517 1.401449 0.016780 9.8661e-01   
factor(STABBR)NC 0.031607 1.166702 0.027091 9.7839e-01   
factor(STABBR)ND 0.043491 1.223148 0.035556 9.7164e-01   
factor(STABBR)NE 0.016017 1.185977 0.013505 9.8923e-01   
factor(STABBR)NH -0.010996 1.433573 -0.007671 9.9388e-01   
factor(STABBR)NJ 0.029751 1.197425 0.024845 9.8018e-01   
factor(STABBR)NM 0.040516 1.581778 0.025614 9.7957e-01   
factor(STABBR)NV 0.004640 1.169946 0.003966 9.9684e-01   
factor(STABBR)NY 0.022173 1.139163 0.019465 9.8447e-01   
factor(STABBR)OH 0.022169 1.156075 0.019176 9.8470e-01   
factor(STABBR)OK 0.026903 1.217696 0.022093 9.8237e-01   
factor(STABBR)OR 0.026461 1.178406 0.022454 9.8209e-01   
factor(STABBR)PA 0.037354 1.144598 0.032635 9.7397e-01   
factor(STABBR)PR 0.004950 1.165247 0.004248 9.9661e-01   
factor(STABBR)SC 0.022766 1.178339 0.019321 9.8459e-01   
factor(STABBR)SD 0.017530 1.262750 0.013882 9.8892e-01   
factor(STABBR)TN 0.031662 1.157814 0.027346 9.7818e-01   
factor(STABBR)TX 0.010694 1.151690 0.009285 9.9259e-01   
factor(STABBR)UT 0.002149 1.274946 0.001686 9.9866e-01   
factor(STABBR)VA 0.025644 1.158194 0.022142 9.8234e-01   
factor(STABBR)VT 0.037158 1.189431 0.031240 9.7508e-01   
factor(STABBR)WA -0.000716 1.249550 -0.000573 9.9954e-01   
factor(STABBR)WI 0.012221 1.210651 0.010094 9.9195e-01   
factor(STABBR)WV 0.029708 1.277122 0.023261 9.8144e-01   
sept2015:highmean -1.331635 0.263069 -5.061922 4.1748e-07 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
RMSE: 8.29877 Adj. R2: 0.065247

## Display result of regression

etable(model1)

model1  
Dependent Var.: index\_sdd\_shmonth  
   
Constant 2.949\*\* (1.145)  
sept2015 -3.23e-6 (0.2061)  
high75 0.3683\* (0.1489)  
public 0.0086 (0.1803)  
private\_np 0.0216 (0.1562)  
factor(year)2014 -2.564\*\*\* (0.1321)  
factor(year)2015 -5.313\*\*\* (0.1383)  
factor(year)2016 -4.973\*\*\* (0.2601)  
factor(STABBR)AL 0.0063 (1.180)  
factor(STABBR)AR 0.0134 (1.204)  
factor(STABBR)AZ 0.0095 (1.296)  
factor(STABBR)CA 0.0123 (1.155)  
factor(STABBR)CO 0.0137 (1.186)  
factor(STABBR)CT 0.0202 (1.186)  
factor(STABBR)DC 0.0024 (1.187)  
factor(STABBR)DE 0.0020 (1.464)  
factor(STABBR)FL 0.0362 (1.159)  
factor(STABBR)GA 0.0264 (1.166)  
factor(STABBR)HI -0.0003 (1.461)  
factor(STABBR)IA 0.0219 (1.167)  
factor(STABBR)ID 0.0225 (1.545)  
factor(STABBR)IL 0.0372 (1.159)  
factor(STABBR)IN 0.0186 (1.181)  
factor(STABBR)KS 0.0492 (1.193)  
factor(STABBR)KY 0.0207 (1.192)  
factor(STABBR)LA 0.0357 (1.163)  
factor(STABBR)MA 0.0296 (1.153)  
factor(STABBR)MD 0.0263 (1.183)  
factor(STABBR)ME 0.0169 (1.212)  
factor(STABBR)MI 0.0281 (1.186)  
factor(STABBR)MN 0.0160 (1.190)  
factor(STABBR)MO 0.0238 (1.160)  
factor(STABBR)MS 0.0330 (1.219)  
factor(STABBR)MT 0.0225 (1.401)  
factor(STABBR)NC 0.0314 (1.167)  
factor(STABBR)ND 0.0435 (1.223)  
factor(STABBR)NE 0.0152 (1.186)  
factor(STABBR)NH -0.0118 (1.435)  
factor(STABBR)NJ 0.0240 (1.195)  
factor(STABBR)NM 0.0405 (1.582)  
factor(STABBR)NV 0.0040 (1.170)  
factor(STABBR)NY 0.0190 (1.138)  
factor(STABBR)OH 0.0205 (1.156)  
factor(STABBR)OK 0.0254 (1.217)  
factor(STABBR)OR 0.0261 (1.178)  
factor(STABBR)PA 0.0357 (1.144)  
factor(STABBR)PR 0.0046 (1.165)  
factor(STABBR)SC 0.0217 (1.178)  
factor(STABBR)SD 0.0160 (1.261)  
factor(STABBR)TN 0.0313 (1.158)  
factor(STABBR)TX 0.0092 (1.151)  
factor(STABBR)UT 0.0011 (1.274)  
factor(STABBR)VA 0.0245 (1.158)  
factor(STABBR)VT 0.0371 (1.189)  
factor(STABBR)WA -0.0027 (1.250)  
factor(STABBR)WI 0.0116 (1.211)  
factor(STABBR)WV 0.0297 (1.277)  
sept2015 x high75 -2.293\*\*\* (0.3295)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedas.-rob.  
Observations 30,285  
R2 0.06789  
Adj. R2 0.06613  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The table above shows the coefficients from estimating model1 in which *high75i* is used as the dummy for high income. The coefficient on the interaction term, -2.293, suggests that the introduction of the College Scorecard decreased search activity on Google Trends for colleges with high-earning graduates by -2.293 standard deviation relative to what it did for colleges with low-earning graduates, with a standard error of 0.3295. This coefficient is significant at the 1% level. State and year fixed effects are controlled for in the regression. Robust standard errors are reported in the parentheses.

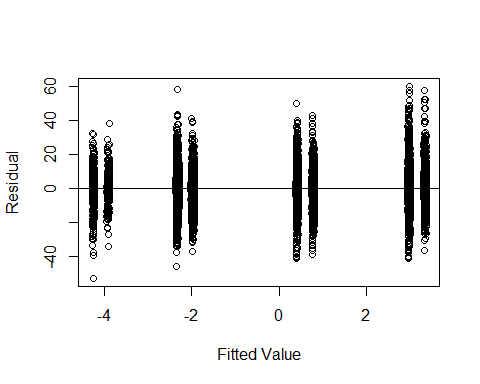
etable(model2)

model2  
Dependent Var.: index\_sdd\_shmonth  
   
Constant 2.950\*\* (1.144)  
sept2015 0.0013 (0.2140)  
highmean 0.2101. (0.1226)  
public 0.0082 (0.1769)  
private\_np 0.0211 (0.1512)  
factor(year)2014 -2.564\*\*\* (0.1321)  
factor(year)2015 -5.313\*\*\* (0.1383)  
factor(year)2016 -4.973\*\*\* (0.2606)  
factor(STABBR)AL 0.0069 (1.180)  
factor(STABBR)AR 0.0138 (1.204)  
factor(STABBR)AZ 0.0117 (1.295)  
factor(STABBR)CA 0.0145 (1.156)  
factor(STABBR)CO 0.0152 (1.186)  
factor(STABBR)CT 0.0212 (1.186)  
factor(STABBR)DC 0.0041 (1.186)  
factor(STABBR)DE 0.0042 (1.465)  
factor(STABBR)FL 0.0374 (1.158)  
factor(STABBR)GA 0.0273 (1.166)  
factor(STABBR)HI 0.0004 (1.467)  
factor(STABBR)IA 0.0228 (1.167)  
factor(STABBR)ID 0.0225 (1.545)  
factor(STABBR)IL 0.0389 (1.159)  
factor(STABBR)IN 0.0200 (1.181)  
factor(STABBR)KS 0.0493 (1.193)  
factor(STABBR)KY 0.0214 (1.193)  
factor(STABBR)LA 0.0368 (1.163)  
factor(STABBR)MA 0.0318 (1.153)  
factor(STABBR)MD 0.0295 (1.186)  
factor(STABBR)ME 0.0170 (1.210)  
factor(STABBR)MI 0.0290 (1.185)  
factor(STABBR)MN 0.0186 (1.192)  
factor(STABBR)MO 0.0241 (1.160)  
factor(STABBR)MS 0.0332 (1.219)  
factor(STABBR)MT 0.0235 (1.401)  
factor(STABBR)NC 0.0316 (1.167)  
factor(STABBR)ND 0.0435 (1.223)  
factor(STABBR)NE 0.0160 (1.186)  
factor(STABBR)NH -0.0110 (1.434)  
factor(STABBR)NJ 0.0297 (1.197)  
factor(STABBR)NM 0.0405 (1.582)  
factor(STABBR)NV 0.0046 (1.170)  
factor(STABBR)NY 0.0222 (1.139)  
factor(STABBR)OH 0.0222 (1.156)  
factor(STABBR)OK 0.0269 (1.218)  
factor(STABBR)OR 0.0265 (1.178)  
factor(STABBR)PA 0.0374 (1.145)  
factor(STABBR)PR 0.0050 (1.165)  
factor(STABBR)SC 0.0228 (1.178)  
factor(STABBR)SD 0.0175 (1.263)  
factor(STABBR)TN 0.0317 (1.158)  
factor(STABBR)TX 0.0107 (1.152)  
factor(STABBR)UT 0.0021 (1.275)  
factor(STABBR)VA 0.0256 (1.158)  
factor(STABBR)VT 0.0372 (1.189)  
factor(STABBR)WA -0.0007 (1.250)  
factor(STABBR)WI 0.0122 (1.211)  
factor(STABBR)WV 0.0297 (1.277)  
sept2015 x highmean -1.332\*\*\* (0.2631)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedas.-rob.  
Observations 30,285  
R2 0.06701  
Adj. R2 0.06525  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*Model2* is examined with *highmeani* being the dummy for high income. Other aspects are help constant as in *model1*. The results are similar. The coefficient on the interaction term is -1.332, suggesting that the introduction of the College Scorecard decreased search activity on Google Trends for colleges with above-average earning graduates by 1.332 standard deviations relative to colleges with below-average earning graduates. This effect is statistically significant at the 1% level. State and year fixed effects are controlled for in the regression. Robust standard errors are reported in the parentheses.

## Draw a graph

res1 <- residuals(model1)  
  
#produce residual vs fitted plot  
plot(fitted(model1), res1, xlab = "Fitted Value", ylab = "Residual")  
  
# add a horizontal line at 0  
abline(0,0)



The graph shows the residual vs. the fitted value of *model1* in which *high75i* is used as the dummy variable for high income. It can be clearly seen that the magnitude of dispersion is not constant as the fitted value increases. For example, the level of dispersion of points is lower when the fitted value is around 4. It is higher when the value is above 2. This shows evidence of heteroskedasticity presented in the data. To solve this problem, I have included robust standard errors in the regressions.

**Conclusion**: Both regressions show consistent negative relationships. Results suggest that the search activity for colleges with high-income graduates had decreased after the release of earning data in Scorecards datasets. Controlling for variables like school type and introducing fixed effects for state and year helps provide a more accurate assessment. These findings contribute to our understanding of the factors influencing students’ college preferences in the context of the College Scorecard release. It will be interesting to dig deeper into this topic and find out the reason behind the negative relationship.