PREDICTIVE ANALYTICS FOR REDUCING STUDENT LOAN DEFAULT

SARAH RILEY

JUNE 2020

UNC COLLEGE OF ARTS AND SCIENCES
Center for Community Capital

Acknowledgements

This report was made possible with generous funding from the Annie E. Casey Foundation. Allison Freeman and Jess Dorrance provided helpful comments on earlier versions of this report. The opinions expressed herein are those of the author.

About the Author

Sarah Riley is a senior research economist at the Center for Community Capital. Her primary research interests are human capital formation and labor market dynamics.

About the Center for Community Capital

The Center for Community Capital is a non-partisan, multi-disciplinary research center housed within the University of North Carolina at Chapel Hill, and is a leading center for research and policy analysis on the power of financial capital to transform households and communities in the United States. The Center's in-depth analyses help policy makers, advocates, and the private sector find sustainable ways to expand economic opportunity to more people, more effectively.

About the Research Program in Higher Education Finance

Policy makers, philanthropists, and researchers have raised questions about how recent trends in student borrowing for higher education may impact our economy and society. The Center for Community Capital's research program in higher education finance seeks to inform public policies and institutional best practices regarding educational debt, student financial literacy, and the future of postsecondary education.

UNC

COLLEGE OF
ARTS AND SCIENCES

Center for
Community
Capital

SUMMARY

Student loan default is largely a symptom of institutional, student, and socioeconomic factors that interact and reinforce each other as students pass through the educational system. Thus, the problem of student loan default is unlikely to be solved at the point of default. Predictive modeling methods provide a potentially powerful means of identifying students at risk of default. Yet the value of model predictions depends crucially on how they are used, and on the quality of the data from which they are generated. A key challenge involves protecting the rights of individuals to decide how their personal information is used while ensuring that predictive models are based on representative data.

INTRODUCTION

Federal student loan debt in the United States exceeds $1.5 trillion, or about 7% of the national Gross Domestic Product.[1] The scale of this debt has generated concern among philanthropists and policy makers, both because the potential exists for student loan debt to deter household formation and displace other types of consumer debt, such as mortgage debt, and because high default rates have been observed among some groups of student borrowers. In light of this concern, this report discusses the strengths and limitations of predictive modeling approaches within the context of their potential applications for reducing student loan default.

Modeling methods for both within-sample and out-of-sample prediction can be used to understand what institutional, student, and socioeconomic factors are associated with student loan default. These methods can also be used to identify high-risk student loan borrowers and design intervention strategies aimed at reducing default rates. Within-sample predictive analysis focuses on explaining or understanding the observed outcomes of an existing group, and it can be used to assess the effectiveness of interventions intended to influence those outcomes. Out-of-sample predictive analysis seeks to forecast the future outcomes of a new group based on previous observations of an existing similar group, and it can be used to inform where interventions of established effectiveness should be targeted in order to influence those outcomes. While the former approach has been widely used in research aimed at identifying the drivers of student loan default, the latter has seen limited application in this context and holds significant potential for identifying and targeting students who are at high risk of defaulting on their student loans.

Because an understanding of the factors associated with student loan default is essential for designing effective interventions aimed at reducing default rates, and because those factors have not changed greatly in the past 50 years, the next section provides an overview of the known predictors of student loan default. These findings, obtained largely from within-sample predictive models intended for statistical inference, suggest broad-brush but potentially high-impact areas for influencing student loan outcomes via the educational pipeline.[2]

1 Author's calculations based on data from the Federal Student Aid Data Center and the Bureau of Economic Analysis.

2 The educational pipeline is commonly defined as the K-12 or K-16 educational system.

The subsequent section digs more deeply into the differences between within-sample and out-of-sample predictive modeling methods, emphasizing the potential of newer technologies to outperform traditional analytic methods for targeting individual students within high-risk borrower groups. The final sections note several opportunities for reducing student loan default via out-of-sample predictive models, as well as some caveats associated with their use.

KNOWN PREDICTORS OF DEFAULT

The current prevalence of student loans for higher education reflects a shift in federal policy that began during the 1970s and 1980s toward the provision of loans for education instead of grants. Policies in the 1990s raised the limits on the amounts that could be borrowed and expanded access to more students. These policy changes and the literature related to student loan default published during the period of 1978-2007 are summarized by Gross et al. (2010), who review 41 studies with emphasis on those using larger sample sizes and multivariate methods. A variety of important findings emerge from their review:

- Program completion is the most important predictor of default across institutions, with default rates historically much higher for non-completers than for completers (14% vs. 2%).
- Students with continuous enrollment who take and complete more credit hours and graduate on time have lower rates of default. Those who graduate on time also tend to borrow less, all else equal.
- Higher standardized test scores prior to enrollment are associated with a lower risk of default, as are higher high-school GPAs, both of which indicate better college preparation. The field or program of study selected by the student also influences default rates via both the amount borrowed and subsequent post-graduate earnings.
- Student personal income following graduation is a key determinant of default, with lower default rates associated with higher incomes and higher default rates associated with unemployment.

- Students perceive student loan debt negatively as the ratio of monthly payment to income increases. Borrowing more overall is associated with a higher risk of default, and a monthly payment to income ratio in excess of 8% is considered to be a burden, except for graduate students, for whom borrowing more is often associated with higher incomes and is inversely related to default.

- Students not attending four-year colleges are more likely to have low incomes, come from racial/ethnic minority groups, and to borrow relatively more in relation to labor market outcomes. Each of these factors is associated with higher default rates and with feelings of being more burdened by debt, as well as with the point of view that educational benefits may not have exceeded the costs.

- Default rates tend to be higher for students with dependents, students who are widowed/separated/divorced, and those who have lower levels of financial support from family members. The likelihood of default also increases with borrower age, all else equal, possibly because of other financial obligations and higher accrued amounts of student debt. Moreover, Black students tend to be more likely to default than White students who have similar labor market outcomes.

- The relationship of grant aid to student loan default is mixed, because receiving grant aid can reduce the cost of college attendance but is also correlated with greater financial need.

- Default rates are not a precise indicator of institutional quality, because they reflect student body characteristics in addition to institutional characteristics.

More recent research generally confirms the continued relevance of these factors for predicting student loan default.[3] In particular, Looney and Yannelis (2015) examine the incidence of federal student loan default during the period of 1999-2014. They find that the increase in default rates observed between 2000 and 2011 can be largely attributed to increases in the rates of college attendance and borrowing among non-traditional borrowers,[4] often from low-income families, who attended proprietary, two-year, or

[3] See Deming, Goldin and Katz (2012), Looney and Yannelis (2015), Looney and Yannelis (2019), Goodell (2016), and Mezza and Sommer (2016).

[4] Non-traditional students tend to be older when they enroll, live independently from their parents, enroll less than full-time, and attend either proprietary or less-than-four-year institutions. The authors also note that these students tend to be disproportionately from low-income households and less affluent neighborhoods and that they are more likely to live in poverty following graduation, as well as less likely to graduate at all. Historically, such borrowers have

other non-selective institutions rather than traditional, selective, four-year public or private (nonprofit) institutions, and who ended up defaulting at higher rates due to poor labor market outcomes. As such, these trends represent growth and change in the composition of the borrowing student population and in the institutions attended compared with earlier periods, with the result that a larger fraction of the federal portfolio represented non-traditional borrowers exiting less selective educational programs and entering repayment in larger numbers during the financial crisis. Because unemployment rates rose much more rapidly for non-traditional borrowers than for traditional borrowers during the recession, these differential labor market outcomes contributed to the higher overall rates of student loan default.

Looney and Yannelis (2015) also note that better oversight of less selective institutions, the decrease in the rate of new borrowing by non-traditional students, and the increase in the use of income-based repayment programs that have taken place since 2010 should ameliorate the trend in student loan default over time. In addition, they point out that the ratio of debt payments to monthly income in recent years has tended to be lowest among non-traditional borrowers at proprietary or two-year institutions and highest among graduate borrowers; although the debt burden relative to income has increased more rapidly at less selective schools in recent years, default rates are actually inversely related to debt burdens.[5] Borrowers with the highest loan balances are more likely to have attended selective four-year institutions or completed graduate study, and they tend to have higher incomes following graduation as well as lower rates of default.

Moreover, Mezza and Sommer (2016) examine a nationally representative sample of student loan borrowers in the US during the period of 1999-2010 and find that credit scores measured just prior to student loan repayment are highly predictive of student loan default during the repayment phase. The authors note that credit scores tend to be positively correlated with socioeconomic background and the likelihood of degree completion, and inversely correlated with the propensity to attend proprietary institutions. This pattern suggests that the higher student loan default rates observed at

represented a small share of the federal student loan portfolio, but they represented a larger share of new borrowers enrolling in higher education just before and during the financial crisis.

[5] For example, in 2011 the median payment to income ratio two years after entering repayment was 5.5% for borrowers at two-year schools and 6.9% for borrowers at proprietary schools, compared with 7.5% for borrowers at selective four-year schools and 9.9% among graduate borrowers.

proprietary institutions partly reflect sorting based on unobserved[6] or socioeconomic factors that influence the ability to repay and are present before loans enter the repayment phase. Mezza and Sommer (2016) also find that student loan balances, which are currently used to determine eligibility for extended or income-based repayment plans, are relatively poor predictors of default, particularly when credit scores are taken into consideration. They suggest that credit scores measured just before the repayment phase may be a better means of identifying borrowers who are most likely to benefit from such repayment plans.

More generally, the tendency of students to repay their student loan debt is highly correlated with their labor market outcomes and with the incomes that they earn after attending college relative to the amount of debt that they incur while in school. An understanding of what determines the returns to higher education, and of how these returns differ across students and institutions, is critical for understanding the factors associated with student loan debt accumulation and default, as well as how optimal educational investment and borrowing decisions vary across individuals. Some key results from the large literature on human capital formation and the returns to education are as follows:[7]

- Labor market outcomes for college students are strongly correlated with student abilities (i.e., cognitive and noncognitive skills), choice of major, degree completion, and the selectivity of the degree-granting institution. Cognitive skills include verbal and mathematical abilities. Noncognitive skills include personality factors, such as conscientiousness and the ability to defer gratification.[8] The standardized test scores frequently used by colleges to screen applicants reflect a combination of cognitive and noncognitive skills.

[6] For example, credit scores are also correlated with personality traits, which are often unobserved in studies of student loan default: https://www.fico.com/blogs/credit-scoring-which-personality-traits-predict-credit-risk

[7] See, for example: Cuhna and Heckman (2007), Hoxby (2009), Riley (2009), Avery and Turner (2012), Heckman, Pinto and Savelyev (2013), Heckman and Mosso (2014), Deming et al. (2016), Webber (2016), Engbom and Moser (2017), Heckman, Humphries and Veramendi (2018), Baker, Gruber and Milligan (2019), and Cornelissen and Dustmann (2019).

[8] Looney and Yannelis (2015) have suggested that personality may also influence student loan default directly.

- Differences in cognitive and noncognitive skills across children from different socioeconomic groups are observable at very early ages and tend to persist over time. To the extent that these ability gaps can be reduced, devoting resources to the early childhood environment represents the most effective avenue for making improvements,[9] although subsequent investments are also required for these improvements to be sustained. Skill investments made at older ages tend to magnify and are magnified by those made at younger ages. For these reasons, an investment in the development and education of a child that begins before primary school and is spread in a balanced way throughout the years of primary and secondary education is more efficient, effective, and equitable than a comparable investment that is made only once the child has reached college age.

- Abilities that are present at (or before) the time of college matriculation are strongly associated with the type and selectivity of college attended, the likelihood of graduation, and the specific skill investments made during college, including choice of major and upper-level coursework in science and mathematics, as well as with subsequent performance on the job. From a labor market perspective, college serves as a screening mechanism for specific abilities, as well as a training program for specific skills, that employers value.

- All else equal, students graduating with degrees in economics, business, and STEM fields (science, technology, engineering, and mathematics) tend to have higher lifetime earnings than graduates majoring in other fields. In general, aptitude for and training in quantitative subjects tends to confer a premium.[10]

- Credentials from proprietary and other less selective institutions are generally viewed less favorably by employers, which may partly explain why graduates from these schools tend to have worse labor market outcomes than those from selective four-year schools. Moreover, the returns to education are mediated by sorting of workers to different types of employers: more selective and better resourced employers tend to achieve a higher concentration of very able and highly educated workers, and the fact that such employers tend to pay higher wages is both a cause and a consequence of this concentration.

---

[9] Noncognitive skills remain malleable for somewhat longer than cognitive skills, but the effectiveness of environmental interventions designed to change either of these skills is lower for older children than for younger ones.

[10] Numeracy is also correlated with financial literacy and better financial behaviors (Lusardi, 2012).

- Overall, college remains an economically beneficial investment for most prospective college students, despite increasing variance in the earnings of college graduates. However, due to the complementarity[11] that exists between ability and education, the returns to college education are highest for high-ability students who graduate from selective four-year institutions, particularly those who study STEM subjects and find work with selective employers. For such students, the expected earnings over the life cycle will generally exceed the cost. In contrast, for low-ability students, particularly those who study arts/humanities and attend proprietary, less-than-four-year, or other non-selective institutions and incur substantial interest-bearing debt in the process, the economic value of a college education is much less clear.

In short, student loan default is largely a symptom of institutional, student, and socioeconomic factors that interact and reinforce each other as students pass through the educational system. Thus, the problem of student loan default is unlikely to be solved at the point of default. Student loan defaults tend to be concentrated among non-traditional students, often from socioeconomically disadvantaged backgrounds, who are less prepared for postsecondary study and who attend less-than-four-year or other non-selective institutions, often do not complete their degrees, and have poor labor market outcomes. Offering this group of postsecondary students career and educational counseling, support services, and/or income-based loan repayment programs represents a potentially high-impact approach to mitigating student loan default risk. This group of students can be identified using information about the institution of attendance, standardized test scores at the time of college entry, and credit scores measured just before the start of loan repayment.

The next section discusses potential ways in which out-of-sample predictive modeling may be used to more precisely focus interventions aimed at reducing student loan default risk among high-risk borrowers. However, as noted above, labor market and student loan repayment outcomes substantially reflect student characteristics that are present ***before*** college entry. Thus, a long-term approach to mitigating student loan default would ideally also involve child savings accounts earmarked for education, combined with investments in early childhood development and ongoing enriched active learning environments that improve college readiness through the development of cognitive and noncognitive skills throughout the entire educational pipeline.

[11] Ability magnifies the value of education, and education magnifies the value of ability.

PREDICTIVE MODELING METHODS & APPLICATIONS

Analytic methods for predictive modeling fall roughly into two major categories, namely (1) traditional statistical or econometric modeling approaches, and (2) newer, more flexible approaches that leverage big data[12] and technological innovations in machine learning (Varian, 2014). There is some overlap between these two groups, and both approaches can be used for out-of-sample prediction and can yield similar results. However, the former methods typically specify a parametric functional form[13] and are used primarily for statistical inference involving within-sample prediction, whereas the latter are often nonparametric and used primarily for identifying generalizable data patterns for the purpose of out-of-sample prediction (Mullainathan and Spiess, 2017; Bzdock, Altman and Krzywinski, 2018).

The table on the next page summarizes the major uses and performance-related considerations underlying these two approaches to predictive modeling. In addition, the following examples illustrate how these methods yield information in the context of student loan default modeling.

Statistical or econometric predictive modeling methods have typically been used by researchers to assess the relative importance of various factors in relation to an outcome of interest. Within the context of student loan debt, these methods have been used to predict default at the institution level and at the student level. For example, Goodell (2016) uses institution-level data to model the default rate as a linear regression function of institution type, location, graduation rate, and student body composition. He finds that proprietary institutions exhibit higher default rates, all else equal.

[12] *Big data* is most commonly defined as data that are high volume, high velocity, and high in variety: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[13] For example, an ordinary least squares (OLS) regression equation assumes that the relationship between the outcome variable and the predictors is linear. The estimated regression coefficients are parameters of this linear model.

METHODS	MOST COMMONLY USED FOR...	PERFORMANCE
Statistics/ econometrics	Hypothesis testing, describing relationships among variables, parameter estimation, understanding what factors cause a particular outcome for an existing sample, and evaluating the effectiveness of interventions.	Model performance improves as sample size increases, but models can perform well with small samples, depending on the level of statistical significance desired. Generally leverages a small number of theoretically important variables. Traditionally used when available data and/or computing power are limited. Unbiased inference and prediction require a representative sample.
Machine learning	Data mining, data classification, and generating out-of-sample predictions that are more accurate and nuanced than those produced by statistical/econometric methods, often without the goal of understanding why a particular algorithm produces the predictions that it does.	Can perform well with moderate sample sizes but is particularly well suited to large samples and a large number of variables (i.e., high-dimensional data). Increasingly used in recent years as computing capabilities and data availability have improved. Unbiased prediction for a rare outcome or subgroup may require an oversample.

Similarly, Looney and Yannelis (2015) use a logistic regression model and student-level data to assess the contribution to default of institution type, pre-enrollment borrower demographic and socioeconomic characteristics, and labor market outcomes among federal student loan borrowers. Using this framework, they determine that borrower type and institution type are statistically significant and substantively important predictors of default, and that the correlation between these factors indicates substantial self-selection (or sorting) by non-traditional borrowers into less selective schools. In particular, they find that borrowers from higher-income households, borrowers who complete more years of school, and borrowers who have higher post-graduation incomes are less likely to default, and that the association of these factors and institution type with repayment outcomes has strengthened over time. While such statistical/econometric models can also be used to generate out-of-sample probabilities regarding whether certain individual student loan borrowers will default in the future, the literature does not currently appear to contain examples of cases in which this has been done.

Newer machine learning methods appear to have received very little application thus far with respect to student loan default.[14] However, the potential of these methods for predictive modeling in this space may be inferred by considering recent findings from analyses of mortgage performance. The logistic regression model, a widely used statistical framework, has traditionally been the workhorse analytic framework used for analyses of mortgage default, and recent attempts have been made to assess the extent to which nonparametric machine learning methods represent an improvement over that traditional approach for accurately predicting mortgage default for individual borrowers.

For example, Fuster et al. (2018) analyze loan-level data for approximately 10 million mortgages originated in the US between 2009 and 2013 and simulate lending decisions (loan approvals and interest rates) for borrowers based on assumptions about whether lenders evaluate borrower default risk using a logistic regression framework or a machine learning framework. They find that loan decisions based on machine learning result in loan approvals that are more inclusive across racial groups than those based on

[14] Preliminary work in developing machine-learning-based predictive models for institution-level student loan default rates using College Scorecard data has been carried out by computer science and mathematics researchers at UNC Greensboro. This work generally confirms the importance of student socioeconomic characteristics and degree completion in predicting student loan repayment: https://arxiv.org/pdf/1805.01586.pdf

logistic regression models, but that they also result in greater dispersion in interest rates, with whites and Asians receiving lower interest rates and Blacks and Hispanics receiving higher interest rates compared with the rates they would have received under the logistic regression framework. These results derive from the fact that the machine learning framework increases the dispersion of default risk estimates for the population compared with the logistic regression framework.

Fuster et al. (2018) interpret their interest rate finding as a negative consequence of the use of machine learning in place of the traditional logistic regression framework, because it results in a reduction in risk pooling ***across*** demographic groups that is achieved through more precise risk-based mortgage pricing and thereby exacerbates intergroup differences in the cost of credit. However, a positive implication of their research is also that machine learning may be better suited than logistic regression to predicting default risk precisely ***within*** groups.

Thus, in the context of predicting student loan default, machine learning represents a potentially powerful means of identifying which individuals from among those categories of borrowers considered to be high risk are most likely to default in practice. Such identification represents an opportunity to ensure that resources aimed at reducing student loan default are allocated with precision.

POINTS OF IMPACT

The existing evidence regarding the factors associated with student loan default suggests several potential postsecondary intervention points at which out-of-sample predictive modeling could be leveraged to influence student loan default risk:

- **Matriculation:** Based on demographic and family characteristics, standardized test scores, and high school academic performance, entering students can be categorized with respect to their likelihoods of taking out student loans, the amounts they are likely to borrow, and their likelihoods of graduating on time (or at all). Supportive advising for first-year students can be tailored accordingly.

- **Major selection:** Students typically declare their major fields during their second year of study. This information, combined with postsecondary grades and other student characteristics, can be used to update the predicted likelihood of graduation for each student, as well as to predict the likelihood of successful loan repayment conditional on likely employment outcomes as implied by the selected field of study. Predictive models can also be developed to forecast the likelihood that students who select less lucrative majors and are also likely to incur significant amounts of student loan debt would be able to complete their degrees in more lucrative majors. Academic advising and/or other support services can be tailored accordingly.
- **Annual or semester enrollment:** Students who enroll part-time are at greater risk of taking longer to graduate or of failing to graduate. These students can end up with relatively higher student loan debt if they graduate (because they have spent more time in school), or with unproductive debt in the event that they do not graduate. Ongoing enrollment information can be used to update the predicted likelihood of graduation, as well as related predictions of the likelihood of student loan repayment in light of likely employment outcomes. Academic advising and/or other support services can be tailored accordingly.
- **Graduation or loan repayment:** Students who complete or discontinue their studies carrying student loan debt can be categorized as to their likelihoods of loan repayment based on their demographic and family characteristics, fields of study, test scores, academic performance, and credit scores. Students can be offered enrollment in income-based loan repayment programs based on this prediction. Advising, preventive loan servicing, and/or other support services can also be tailored accordingly.

Such models could be leveraged more broadly by creating or licensing a data-driven software application that would permit students to enter their personal characteristics (demographics, grades, and test scores) in combination with different possible colleges, majors, and job roles. This tool would enable students to assess the likely costs and benefits of different educational investment decisions at different points in their postsecondary careers based on the recent experiences of similar students. Such an application could be used as an adjunct to conventional college counseling at the high-school level and could be leveraged directly by postsecondary students at matriculation, when choosing a major, during annual enrollment, when applying for internships, and at graduation. Employers routinely use a variety of individual and market data in predictive models that inform their hiring processes. Putting similar information and tools in the hands of students and their families could help to facilitate more informed decision-making across multiple dimensions of the postsecondary experience, including but not limited to student loan debt and repayment.

CAVEATS

Despite the potential of new technologies and data to revolutionize the management of student loan debt and default risk, caution is warranted in several respects:

- Predicting which students are most likely to default is not necessarily the same thing as predicting which students are likely ***both*** to default ***and*** to respond to an intervention intended to reduce default.[15] Interventions may be more effective for some borrowers than for others, and some interventions may generally be ineffective. Thus, it may be advisable to conduct controlled experiments to assess the effectiveness of certain types of interventions for different student groups before offering those interventions to students.
- Although having accurate information regarding their likely personal costs and benefits from a college education may help students to make more economically efficient education and career decisions, it may also discourage some students from pursuing higher education or aspirational career goals. To the extent that realism helps some students to avoid unproductive debt, such an outcome may be desired. However, it may be best to provide this information within the context of supportive career counseling. While the insight derived from predictive models may be useful for making informed decisions, it should not be treated as a limiting and deterministic forecast of what is possible for any particular student. More generally, the value of model predictions will depend on how those predictions are used.

[15] For further discussion of this issue, see Athey (2017).

- Predictions generated by predictive models are only as good as the data on which those predictions are based. Accurate out-of-sample predictions require that the sample for which predictions are being generated is similar to that on which the model was trained, that the data are correct, and that subgroups of interest are adequately represented. Thus, data quality is a key concern whenever predictive modeling is applied.
- New methods of predictive modeling typically leverage high-dimensional data, which raises concerns about individual privacy and consent. It is difficult or impossible to completely de-identify high-dimensional data, as current algorithms can often accurately infer a person's identity from data that do not include personal identifiers. Moreover, although analyses of personal information conducted in academic and governmental research settings are governed by ethical guidelines that are designed to protect human subjects, such restrictions often do not govern the use of personal data by other organizations or for other purposes. In addition, it is becoming increasingly difficult to participate in educational and employment opportunities without providing personal data and agreeing to their use for analytics. Thus, a key challenge involves protecting the rights of individuals to decide how their personal information is used while ensuring that predictive models are based on representative data.

REFERENCES

Athey, Susan. 2017. "Beyond Prediction: Using Big Data for Policy Problems." *Science*, 355: 483–485.

Avery, Christopher, and Sarah Turner. 2012. "Student Loans: Do College Students Borrow Too Much – Or Not Enough?" *Journal of Economic Perspectives*, 26(1): 165–192.

Baker, Michael, Jonathan Gruber, and Kevin Milligan. 2019. "The Long-Run Impact of a Universal Child Care Program." *American Economic Journal: Economic Policy*, 11(3): 1–26.

Bzdock, Danilo, Naomi Altman, and Martin Krzywinski. 2018. "Statistics versus Machine Learning." *Nature Methods*, 15(4): 233–234.

Cornelissen, Thomas, and Christian Dustmann. 2019. "Early School Exposure, Test Scores, and Noncognitive Outcomes." *American Economic Journal: Economic Policy*, 11(2): 35–63.

Cuhna, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Association Papers and Proceedings*, 97(2): 31–47.

Deming, David J., Claudia Goldin, and Lawrence F. Katz. 2012. "The For-Profit Postsecondary School Sector: Nimble Critters or Agile Predators?" *Journal of Economic Perspectives*, 26(1): 139–164.

Deming, David J., Noam Yuchtman, Amira Abulafi, Claudia Goldin, , and Lawrence F. Katz. 2016. "The Value of Postsecondary Credentials in the Labor Market: An Experimental Study." *American Economic Review*, 106(3): 778–806.

Engbom, Niklas, and Christian Moser. 2017. "Returns to Education through Access to Higher-Paying Firms: Evidence from US Employer-Employee Data." *American Economic Review: Papers and Proceedings*, 107(5): 374–378.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2018. "Predictably Unequal? The Effects of Machine Learning on Credit Markets." Working Paper.

Goodell, John W. 2016. "Do For-Profit Universities Induce Bad Student Loans?" *The Quarterly Review of Economics and Finance*, 61: 173–184.

Gross, Jacob P. K., Osman Cekic, Don Hossler, and Nick Hillman. 2010. “What Matters in Student Loan Default: A Review of the Research Literature.” *Journal of Student Financial Aid*, 39(1): 19–29.

Heckman, James J., and Stefano Mosso. 2014. “The Economics of Human Development and Social Mobility.” *Annual Review of Economics*, 6: 689–733.

Heckman, James J., John Eric Humphries, and Gregory Veramendi. 2018. “Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking.” *Journal of Political Economy*, 126(S1): S197–S246.

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. “Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes.” *American Economic Review*, 103(6): 2052–2086.

Hoxby, Caroline M. 2009. “The Changing Selectivity of American Colleges.” National Bureau of Economic Research Working Paper No. 15446.

Looney, Adam, and Constantine Yannelis. 2015. “A Crisis in Student Loans? How Changes in the Characteristics of Borrowers and the Institutions They Attended Contributed to Rising Loan Defaults.” *Brookings Papers on Economic Activity*, Fall: 1–89.

Looney, Adam, and Constantine Yannelis. 2019. “How Useful Are Default Rates? Borrowers with Large Balances and Student Loan Repayment.” *Economics of Education Review*, 71: 135–145.

Lusardi, Annamaria. 2012. “Numeracy, Financial Literacy, and Financial Decision-Making.” National Bureau of Economic Research Working Paper No. 17821.

Mezza, Alvaro, and Kamila Sommer. 2016. “A Trillion-Dollar Question: What Predicts Student Loan Delinquencies?” *Journal of Student Financial Aid*, 46(3): 16–54.

Mullainathan, Sendhil, and Jann Spiess. 2017. “Machine Learning: An Applied Econometric Approach.” *Journal of Economic Perspectives*, 31(2): 87–106.

Riley, Sarah F. 2009. “Renaissance Man Meets the Pin Factory: A Theory of Diverse Specialization Under Uncertainty.” Department of Economics, University of North Carolina at Chapel Hill Dissertation.

Varian, Hal R. 2014. “Big Data: New Tricks for Econometrics.” *Journal of Economic Perspectives*, 28(2): 3–28.

Webber, Douglas A. 2016. "Are College Costs Worth It? How Ability, Major, and Debt Affect the Returns to Schooling." *Economics of Education Review*, 53: 296–310.

UNC

COLLEGE OF
ARTS AND SCIENCES

Center for
Community
Capital