The Long Journey Through Student Loan Repayment

A look at diverse pathways

Erin Dunlop Velez, Austin Lacy, Michael Duprey,
Johnathan Conzelmann, and Nichole Smith
RTI International

Overview

Many students take out loans to be able to afford a college education. In 2016, some 70 percent of U.S. undergraduates in their fourth year of college or above had taken out loans at some point in their college careers, borrowing an average of $29,000 during their postsecondary education.[1] After leaving school, by completing their degrees or dropping out, or after dropping below half-time enrollment, students begin to repay their loans.

The standard repayment plan schedules monthly payments so that borrowers will pay off the principal and interest on their loans within 10 years if payments are made in full and on time. However, many borrowers experience repayment difficulties. Twelve years after beginning college in the 2003-04 school year, a fifth of federal borrowers (21 percent) had used a deferment for economic hardship.[2] Over a quarter of federal borrowers (28 percent) had defaulted on their student loans. Other borrowers had experienced additional repayment difficulties, such as delinquencies (missed loan payments).

Due to the wide range of repayment difficulties, borrowers often do not take the direct pathway through repayment of making payments each month until their balance is paid off. This report describes the diverse and potentially difficult pathways borrowers take. Key findings include:

- Over half of borrowers experienced negative amortization (loan balances increasing over time because payments are less than the interest accrued).
- Almost half of borrowers exhibited characteristics associated with repayment distress, such as default or an economic hardship deferment.[3]
- Only about a third of borrowers followed the traditional repayment pathway of paying down their balances without experiencing distress, such as default or an economic hardship deferment.

About repayment

When financing their postsecondary education, students can borrow from the federal government or private entities such as banks and credit unions. Compared to private student loans, federal loans offer lower interest rates and more protections for borrowers who run into repayment difficulties. Of all undergraduates in 2016, only 5 percent used private student loans.[4] This report focuses exclusively on repayment of federal student loans.

In general, once students exit postsecondary education or drop below half-time enrollment, their federal student loans are in a grace period for six months, meaning no payments are due.[5] After the grace period ends, payments are due each month until the loans are paid off unless a

borrower re-enrolls in college, either to finish a degree or earn an additional credential, or increases enrollment to at least half-time, in which cases payments on federal student loans are no longer required under an education deferment.

If borrowers in repayment run into difficulty making scheduled loan payments—for example, if they lose their jobs—they can apply for an economic hardship deferment. This deferment temporarily stops required payments on the borrower's loans, but for many loans interest continues to accrue. Another option for borrowers struggling to afford their loan payments is to enroll in a repayment plan other than the standard 10-year plan. For instance, there are several Income-Driven Repayment (IDR) plans, in which payments are set not to exceed a certain threshold of borrower income.[6] Borrowers who do not repay or take advantage of deferments or IDR plans may end up with loans in default. Federal student loans enter default when a borrower makes no payment for 270 days.[7] Once a student defaults, debt collectors may be utilized; loan payments may be collected through nonvoluntary means, such as wage garnishment; and penalties and fees accrue. Defaulted federal student loans are very rarely discharged, even in bankruptcy.

About the data

The borrowers analyzed in this report came from a nationally representative sample of first-time postsecondary students who began college in the 2003-04 school year. Their loan repayments were examined over the subsequent 12 years. The number of semesters a student was enrolled over the 12-year period determined the number of quarters of repayment observed for each borrower. Borrowers enter repayment six months after their enrollment ends. So if a borrower had been enrolled for four years, for example, he or she entered repayment 4.5 years after first enrolling, and we followed the borrower's repayment for the remaining 7.5 years. For consistency and to minimize the truncation of borrowers' loan histories, we limited the sample to borrowers for whom we could observe at least three years (12 quarters) of repayment.

Variation in the number of semesters enrolled, and in the time borrowers took to repay their loans, contributed to the fact that for some borrowers, we observed their complete repayment history (i.e., until their loans were paid off), while for others, we observed only part of their repayment histories. For borrowers still repaying 12 years after beginning postsecondary education, their paths, which are described in this report, are a snapshot of where they were in the repayment process at that time.

This analysis used a modeling approach called a hidden Markov model (HMM) to uncover the pathways borrowers pass through during repayment. The model estimated different possible

repayment *states* that describe the borrower's loan repayment status. Borrowers move through these states in different orders. Each combination of states describes the *pathway* a borrower takes from entering repayment to paying off his or her loans.

Including students paying off their debt completely, the model calculated five distinct states that borrowers inhabit. Figure A.1 describes the repayment behaviors exhibited by borrowers in each state. The specific states are:

- **State 1.** Most borrowers have not started making payments, or they have started but still owe more than they originally borrowed.
- **State 2.** Most borrowers are repaying but still owe more than half of what they originally borrowed.
- **State 3.** Most borrowers are in distress—either in default or in economic hardship deferments.[8]
- **State 4.** Most borrowers are repaying and owe less than half of what they originally borrowed.
- **State 5.** Borrowers have completed repaying their loans.

Figure A.2 describes common movements between the five states. The probabilities in Figure A.2 are the likelihood a borrower moves from one state to another in any given quarter. Notice that the transition probability to stay in state 3, the distressed state, is particularly high, indicating that once students are in distress on their student loans, it often takes them a long time to recover.

The different combinations of states that borrowers move through are the different pathways to repayment completion. Table A.1 lists the proportion of borrowers who took each pathway. Different types of students take different pathways through repayment. We grouped the many pathways into two categories: pathways in which borrowers experienced the distressed state and pathways in which they did not. Table A.2 summarizes the characteristics of borrowers within each group of pathways.

Key terms

- **Default.** Federal student loans go into default after a borrower makes no payment for 270 days.
- **Economic hardship deferment.** If borrowers are struggling to afford the payments on their student loans, they can often work with their servicer to obtain a deferment. When loans are in deferment, no payments are due, so the borrower won't default. Still,

for many loans, interest continues to accrue, meaning the borrower's overall loan balance increases.

- **Negative amortization.** This is when the balance on a borrower's loans increases over time because the borrower makes no payments or makes payments that do not keep up with the interest that accrues.
- **Repayment *states.*** The HMM model uses five unique repayment states that describe borrowers' repayment statuses over time.
- **Repayment *pathways.*** Each borrower may move through the repayment states in a different order. The unique combination of repayment states a borrower moves through between entering repayment and paying off the loan is that borrower's repayment pathway.

Key findings

Over half of borrowers experienced negative amortization

Some 61 percent of borrowers had loan balances that increased between two consecutive quarters. While some of these loans were in education deferments, with increasing balances because borrowers were not making payments while re-enrolled in school, many were not. Over half of borrowers (55 percent) had outstanding principal loan balances increase between quarters while their loans were not in education deferments. This is the definition of negative amortization used throughout the remainder of this report.[9]

While having one's loan balance increase between quarters for any length of time can slow a borrower's progress through repayment, longer periods of negative amortization can have more lasting effects. Some 24 percent of borrowers had outstanding principal balances that increased over three consecutive quarters, while 8 percent of borrowers had balances that increased over four consecutive quarters.

Although periods of negative amortization often lead to students passing through the distressed state (characterized by default or economic hardship deferments), some borrowers corrected their repayment trajectory. Of borrowers who experienced negative amortization, about two-thirds (70 percent) also experienced the distressed state, while the rest did not.

Compared to borrowers who did not experience negative amortization, those who did were:

- Older and more likely to be independent students.[10]
- From lower-income families.
- More likely to be black and less likely to be white.

- More likely to have begun their education at a for-profit institution and less likely to have begun at a public or private nonprofit institution.
- More likely not to be enrolled in a degree program and less likely to be enrolled in an associate degree program.
- More likely to have dropped out or earned a certificate and less likely to have earned a bachelor's degree.
- More likely to have borrowed in the highest quintile of total loan amount—more than $20,500—and taken out a larger number of individual loans.

Almost half of borrowers exhibited characteristics associated with repayment distress, such as default or an economic hardship deferment

Some 48 percent of borrowers encountered the distressed state, in which most borrowers were in default or economic hardship deferments. About 80 percent of those who experienced the distressed state went through periods of negative amortization beforehand.

Compared to borrowers who did not experience the distressed state, those who did:

- Were older and more likely to be independent students.
- Were from lower-income families.
- Were more likely to be black or Hispanic, and less likely to be white.
- Were more likely to have begun their education at a for-profit institution, and less likely to have begun at a public or private nonprofit institution.
- Were more likely not to be enrolled in a degree or certificate program, and less likely to be enrolled in an associate degree program.
- Earned a lower GPA when last enrolled.
- Were more likely to have dropped out or earned a certificate and less likely to have earned a bachelor's degree.
- Were less likely to have borrowed in the highest two quintiles of total loan amount—more than $12,200.

Only about a third of borrowers followed a traditional repayment pathway, paying down their balances without experiencing distress

Some 52 percent of borrowers never passed through the distressed state. But even many of these borrowers experienced negative amortization, with their loan balances increasing. About a third (31 percent) of borrowers who never passed through the distressed state experienced negative amortization. This indicates that only 36 percent of all borrowers experienced a

traditional pathway—never passing through the distressed state and never experiencing negative amortization.

Of those who appear to be traditional borrowers in our sample, however, we observed only two-thirds (69 percent) completing repayment, meaning they went all the way through repayment in the traditional pathway. The other third of traditional borrowers did not complete repayment during our study period. As such, the proportion of our sample that completes repayment in the traditional pathway will be between 25 and 36 percent.

This analysis of pathways indicates that there are early warning signs before borrowers go into distress. The periods of negative amortization before loan repayment distress suggest that these borrowers could be identified and provided with resources early in the repayment process.

Time to repayment

We observed borrowers for up to 12 years after they first enrolled in postsecondary education, but because many students were enrolled for much of this period, we did not observe their repayment completion. Of the students we observed in repayment for at least 10 years, 60 percent reached state 5 and completed loan repayment. For the sample as a whole, however, we observed only 46 percent finish repayment. For that 46 percent, the number of quarters it took borrowers to complete repayment varied widely, from zero to 48 quarters. The average time to complete repayment was 18 quarters (4.5 years).

Had we observed borrowers for a longer period, the range over which borrowers completed repayment would have been even wider. Among an older cohort of students who began college in 1995-96, the average time between entering repayment and paying off all federal student loans was seven years, and 30 percent of those borrowers took more than 10 years (the standard repayment window).

Conclusion

Student experiences with repayment, and the issues students face during repayment, vary. We documented 58 pathways through repayment, which can be characterized by different combinations of five unique states. (See Table A1.) While some students experienced a traditional pathway through repayment, paying down their balances without experiencing distress such as default or an economic hardship deferment, many others had periods of distress. There was also considerable variation in the timing of repayment difficulties. Some

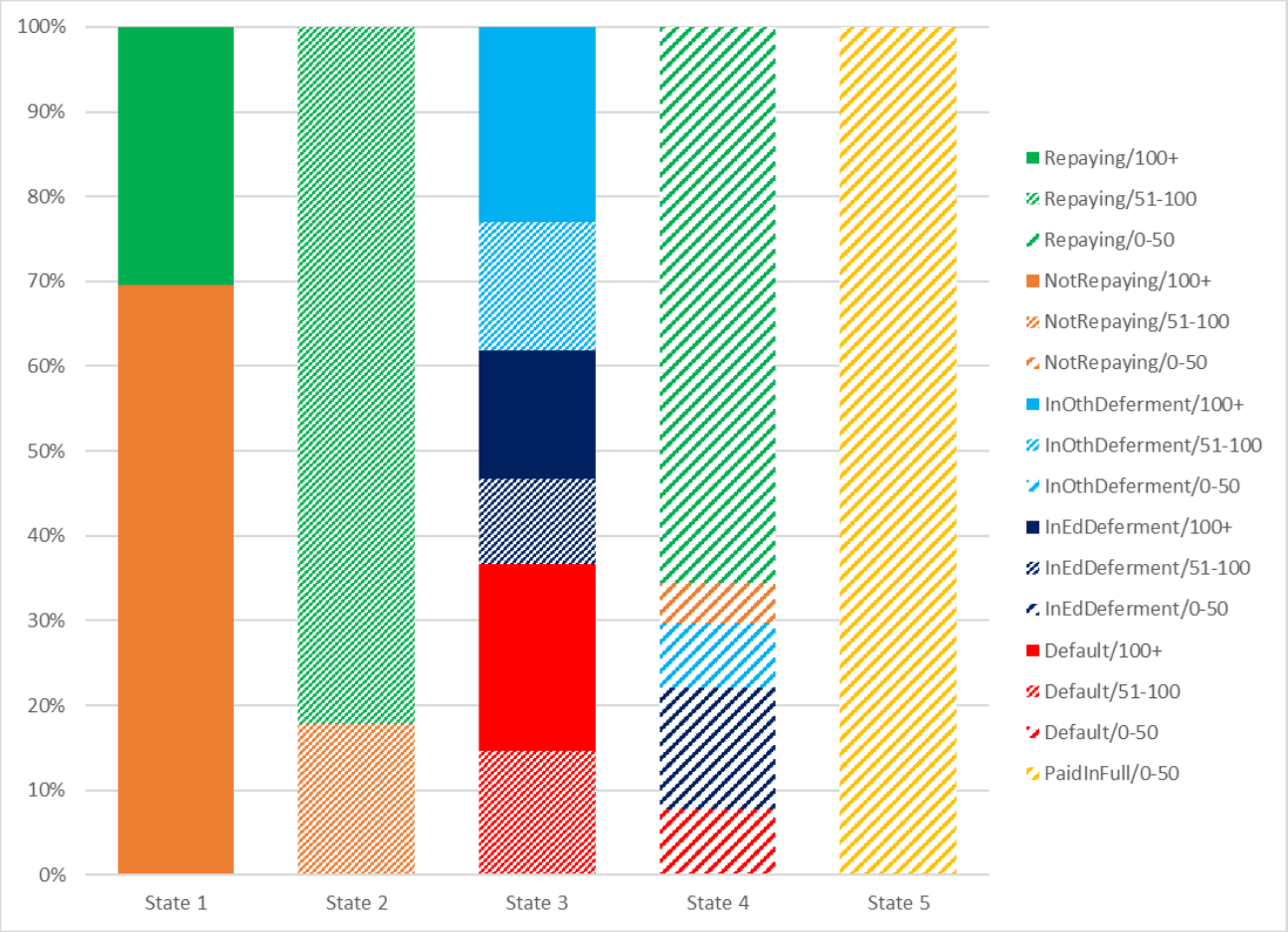
students ran into issues soon after they entered repayment, while others had paid down significant portions of their debt before encountering problems.

Finally, there was also large variation in the amount of time it took borrowers to complete repayment, with some borrowers paying off their loans nearly as soon as they entered repayment, while others took longer than the 10-year standard repayment window. Given the diversity of student pathways through repayment and the variation in the time it takes students to complete repayment, targeted policy approaches will be needed to help those borrowers struggling the most.

Appendix A: Figures and Tables

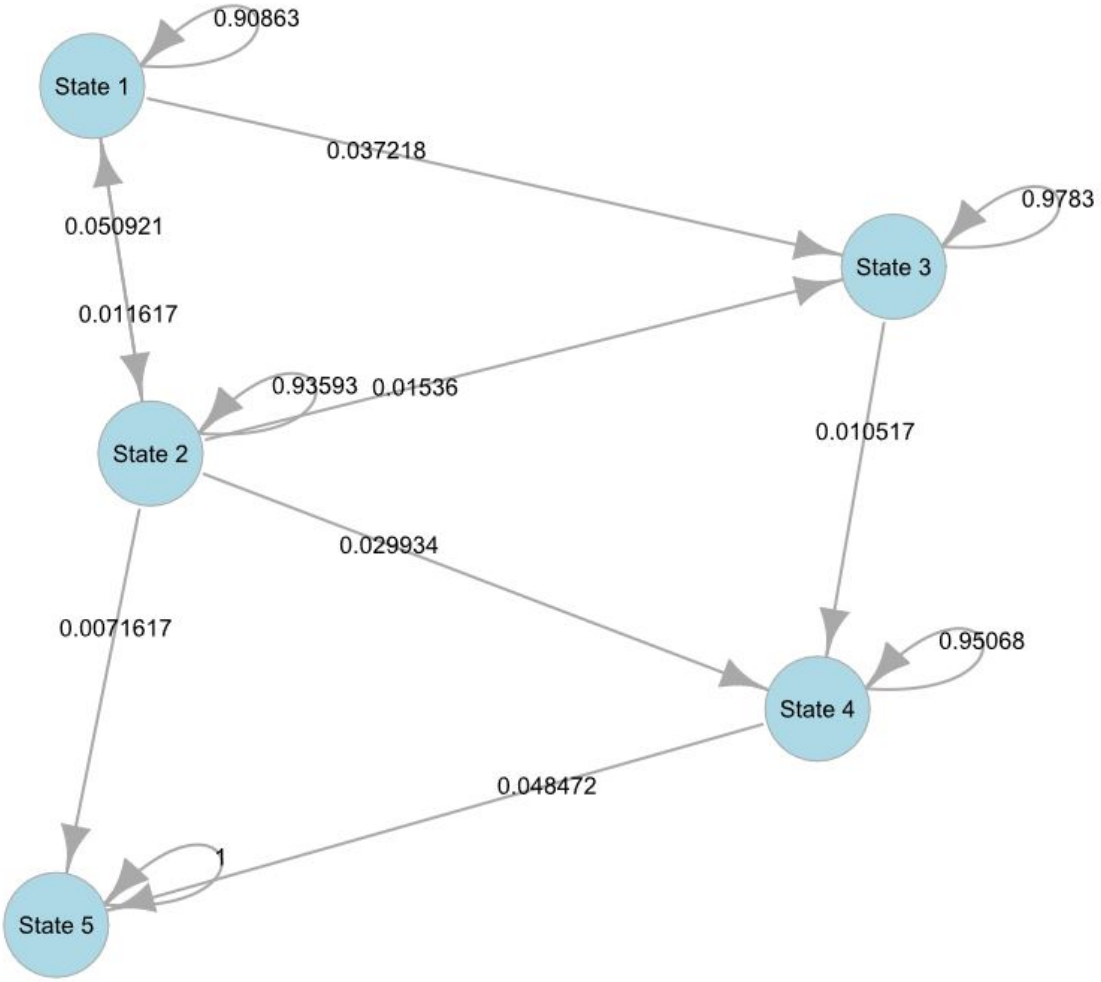
Figure A.1

Composition of Borrowers in Each Repayment State



Note: Each state is composed of different combinations of the borrower's loan status—repaying, not repaying, in deferment other than education deferment, in education deferment, in default, and loans paid off in full—and the percentage of the borrower's loan balance that is paid off. Borrowers may owe more than 100 percent of their original balances due to accrued and capitalized interest.

Figure A.2
Transition Probabilities Each Quarter Between States



Note: The probabilities shown are the likelihood that a borrower moves between any two states in a given quarter. Only probabilities greater than 0.5 percent are shown.

Table A.1

Percentage of Borrowers in Each Pathway

Pathway	Percentage of borrowers
1-3	11.76
2-4-5	11.68
1-2	6.35
1-2-4-5	5.55
1-2-4	4.81
2-4	4.39
2-5	4.24
3	3.95
2-1-3	3.79
2-3-4-5	3.64
2-3	3.31
1	3.25
1-2-3	3.02
1-2-5	2.17
5	2.01
2-3-5	1.95
1-3-4-5	1.73
4-5	1.53
1-3-5	1.40
2-1-3-5	1.36
2-1-3-4-5	1.30
3-4-5	1.27
3-5	1.23
2	1.23
2-1	1.20
2-3-4	1.18
1-3-4	1.08
2-1-4-5	1.08

Note: The table includes the 28 pathways with at least 1 percent of borrowers. There are another 30 pathways with less than 1 percent of borrowers in each.

Table A.2

Characteristics of Borrowers in Each Pathway

	Experienced distressed state			Did not experience distressed state		
	Total	**Had negative amortization**	**No negative amortization**	**Total**	**Had negative amortization**	**No negative amortization**
Ages 18 or younger	0.38	0.34	0.57	0.52	0.48	0.54
Age 19	0.22	0.21	0.26	0.26	0.20	0.28
Ages 20-23	0.17	0.19	0.09	0.10	0.13	0.08
Ages 24-29	0.12	0.14	0.02	0.05	0.08	0.04
Age 30 or older	0.11	0.12	0.06	0.08	0.12	0.06
Independent	0.36	0.41	0.12	0.18	0.27	0.13
Female	0.59	0.60	0.54	0.55	0.55	0.54
White	0.52	0.49	0.66	0.72	0.60	0.77
Black/African American	0.22	0.26	0.10	0.06	0.13	0.03
Hispanic/Latino	0.17	0.18	0.13	0.13	0.17	0.11
Other race	0.09	0.08	0.12	0.09	0.10	0.09
Zero expected family contribution	0.35	0.40	0.15	0.17	0.25	0.13
Family income before college	$40,830	$35,718	$61,845	$60,483	$49,116	$65,658
Parents' education: No college or unknown	0.46	0.50	0.28	0.34	0.43	0.30
Parents' education: Some college	0.27	0.26	0.31	0.26	0.25	0.26
Parents' education: Bachelor's or higher	0.27	0.24	0.41	0.40	0.33	0.44
First institution: Public	0.53	0.49	0.68	0.67	0.63	0.69
First institution: Private not-for-profit	0.13	0.10	0.24	0.17	0.13	0.20
First institution: Private for-profit	0.34	0.40	0.09	0.16	0.24	0.12
First program: Classes only	0.23	0.28	0.06	0.11	0.17	0.08
First program: Certificate	0.40	0.41	0.34	0.35	0.38	0.33
First program: Associate	0.32	0.27	0.52	0.48	0.38	0.52
First program: Bachelor's	0.05	0.05	0.07	0.06	0.07	0.06
Most recent GPA	2.99	2.97	3.05	3.20	3.15	3.22
After 6 years: No degree	0.57	0.59	0.47	0.38	0.48	0.34

After 6 years: Certificate	0.14	0.16	0.07	0.10	0.13	0.09
After 6 years: Associate	0.10	0.10	0.10	0.09	0.08	0.09
After 6 years: Bachelor's	0.19	0.15	0.36	0.43	0.30	0.48
Amount borrowed: Quintile 1	0.23	0.22	0.28	0.22	0.19	0.24
Amount borrowed: Quintile 2	0.26	0.28	0.17	0.19	0.17	0.20
Amount borrowed: Quintile 3	0.19	0.18	0.23	0.16	0.16	0.16
Amount borrowed: Quintile 4	0.15	0.13	0.22	0.22	0.21	0.23
Amount borrowed: Quintile 5	0.17	0.18	0.11	0.20	0.28	0.17
Number of loans	4.26	4.43	3.57	4.21	4.83	3.93

Notes: Individual pathways were combined into groups of pathways based on whether the pathway passed through the distressed state and whether the borrower experienced negative amortization. Students are considered independent for financial aid purposes if they are 24 years of age or older, are married, have children, or are veterans of the military. Quintile 1 of amount borrowed is $1-$3,400. Quintile 2 is $3,401-$6,600. Quintile 3 is $6,601-$12,200. Quintile 4 is $12,201-$20,500. Quintile 5 is $20,501-$226,100.

Appendix B: Methods

About the data

The data used in this report are from the 2015 Federal Student Aid Supplement for the 2004 Beginning Postsecondary Students Longitudinal Study Cohort (BPS). The 2015 aid supplement was created by the U.S. Department of Education's National Center for Education Statistics using two cohorts of students who began college for the first time in 1996 and in 2004. Before the 2015 aid supplement, each cohort had been surveyed three times: during the year they began college, three years later, and then six years after starting college. To create the 2015 aid supplement, the same students were subsequently matched to the National Student Loan Data System (NSLDS) in 2016, and the new administrative data from NSLDS were added and rereleased with the original study data for the two cohorts in 2017.[11] The new data enable researchers to conduct dynamic analyses of long-term federal student loan borrowing and repayment outcomes in a way that is not possible with any other publicly available data set to date. For this study, we focused only on the more recent BPS:04 cohort because outstanding student loan balance histories were not stored in NSLDS until 2005, prohibiting us from tracking repayment for a large portion of the earlier BPS:96 cohort.

Our analytic sample of interest was thus students who began college in the 2003-04 academic year, borrowed federal student loans, and entered repayment on those loans within 12 years of beginning college.[12] To analyze the repayment patterns of this population, we conducted extensive data management of the source files that accompany the 2015 aid supplement and converted the data into person-level histories of federal student loan repayment. This included reshaping administrative records of loan originations, disbursements, maturity dates, outstanding balance and interest histories, deferment periods, default occurrences, and repayment plan information.[13] Once organized into a quarterly data set, we merged our history file with student-level variables provided on the derived data set for the BPS:04 cohort.[14] Though time-intensive, converting the source files to a quarterly data set was necessary for the tracking of repayment.

In its simpler form, the BPS data set represents a group of students who began college at the same time; however, it is not a group of individuals who entered repayment together. That is, students can depart postsecondary education (by dropping out or graduating) at different times and thus enter repayment at different times. To address this, we standardized time in repayment by introducing each borrower to the data set during the quarter when he or she first entered repayment on any federal student loan. This, by definition, reduced the initial BPS sample to students who borrowed federal student loans and who entered repayment on at least one of those loans within 12 years of beginning postsecondary education. To simplify our

analysis, we reduced the sample further to students who entered repayment only once, defined as having all loans entering repayment within two quarters (six months) of one another. Finally, we limited the sample to those borrowers for whom we could have observed at least 12 quarters in repayment to exclude any borrowers whose repayment histories were severely truncated due to data constraints. In our final analytical data set (*n borrowers* ~ 5,600), we observed borrowers for between one quarter (if they paid off all their loans within the first quarter or earlier) and 50 quarters (12.5 years, the maximum possible in the data), depending on how long students were enrolled in postsecondary education. A small number of students (*n* ~ 70) entered repayment at a time such that their repayment patterns should have been observed for 12 quarters, but the data were missing for their final quarters (median of two quarters missing). For these 70 cases, we imputed their missing repayment data by carrying forward their last reported balance and repayment status.

About the method

High-level description

As our research question was concerned with the *pathways* a student makes through repayment—in particular, pathways leading to full repayment—we used as our class of models the *hidden Markov model* (HMM). In a Markov model, entities traverse a series of discrete *states* over time in which the current state position is conditionally dependent only on the most recent previous state.

In a traditional (non-hidden) Markov model these states are directly observable, thus the model may be encapsulated by a matrix containing the probability of each pairwise state transitions and a vector representing initial state position. Under this model, we assume the process is Markovian. However, in our case, we use an unobserved (hidden) layer of loan repayment states through which borrowers pass over time. Position in this hidden layer varies through time as with an observable Markov model, though state positions can only be inferred though a second layer of data we can observe. In HMM terms, the hidden state *emits* observable attributes over some probability distribution. We use two separate data sequences, which each vary by quarter (our unit of interest), or channels, to represent borrowers' observable attributes: (1) loan statuses and (2) outstanding principal balance (OPB) percentages.

Given our two channels of observable repayment attributes, our desire was to estimate the parameters governing the linkage between the hidden states and observable attributes as well as the transition matrix between hidden states. Examination of these parameters and recovery of the hidden states can then allow us to answer questions regarding the types of pathways borrowers take in the process of repaying their student loans.

Limited horizon assumption

Implicit in the Markov model is a notion of the limited horizon assumption, in which states with time t contain enough information to predict time $t + 1$. Or, more formally,

$$P(z_t \mid z_{t-1}, z_{t-2}, \dots, z_1) = P(z_t \mid z_{t-1}),$$

where z_t represents some unobserved (i.e., hidden) state z at time t. To verify the validity of this assumption given our data set, we examined the subset of students for whom we observed 12 or more quarters of repayment. Looking at the first 11 quarters, we examined all of the potential patterns of observed channel 2 statuses, which resulted in over 950 different sequences for our data set. When subsetting the sequences to the dichotomous indicator of those who paid in quarters 1-11, and those who did not pay in quarters 1-10 but paid in quarter 11, we found that in both groups, over 85 percent paid in quarter 12, though those who had repaid consistently did show a higher rate. We also explored this in a probabilistic setting by constructing multinomial logistic regression models using the status at quarter 11 to predict the status in quarter 12, including a separate category for those who paid all 11 quarters. We found a similar sign and statistical significance between those individuals who paid all quarters and those who did not pay all quarters but did repay in quarter 11, though with a larger magnitude for those paying all quarters. Given these observations, we believe the limited horizon assumption holds.

Channels

We restricted our analyses to two channels only for modeling simplicity and ease of interpretability. While HMMs can accommodate an arbitrary number of channels, convergence issues may arise if the channels are numerous, have too many levels within each channel, or encode attributes that are generatively dissimilar from one another. Furthermore, if the primary modeling goal is *interpretation* rather than strict *prediction*, channels are best kept to a small number to aid the researcher in developing an intuitive sense of each state's multivariate composition. In this case, we found that two was the optimal number of channels for interpretation, as the number of levels multiplicatively increases with each additional channel.

The first channel describes an individual student's repayment status at time point t. Repayment statuses are mutually independent and represent whether a student was in repayment, in education-related deferment, in other deferment, not currently repaying, had paid the loans in full, or had defaulted on the loans at time point t, represented by quarters, with a maximum number of 50 quarters available. While the data represent students who began their postsecondary education together, not all students entered repayment at the same time;

therefore, a substantial proportion of the available sample was right censored, meaning terminal repayment behavior was unobserved for such students. Thus, our model represents only students' transitions over the first 12.5 years (50 quarters) of repayment and should not be interpreted as a complete representation of repayment patterns over the entire life of a loan.

The second channel contains a stream of each student's OPB remaining at time point t, as a proportion of cumulative amount of loans disbursed. While a legitimate design choice would be to use total amount owed at time point t instead, we choose to use the percent of OPB still owed because it contains more information about a borrower's progress through repayment. For example, a sample member who initially borrowed $20,000 and later was observed owing $40,000 would have a percent OPB of 200 percent. On the other hand, a borrower who also owes $40,000 but initially borrowed $100,000 would have a percent OPB of 40 percent. From this we have theoretical reasons to believe that percent OPB will better aid us in uncovering our hidden states. For modeling simplicity and to allow for discrete combinations of levels representing repayment status and percent OBP remaining, these proportions were trichotomized into intervals representing 0-50 percent, 51-100 percent, and 100-plus percent remaining. Note that students could accumulate OPBs of greater than 100 percent due to negative amortization in which payments are not large enough to cover the interest due; we allowed the model to incorporate this important status by representing negative amortization that surpasses the original loan balance as a distinct level (i.e., 100-plus percent remaining). While we considered incorporating additional time-varying channels into the HMM, such as scheduled payment amount or actual payment amount, increased model complexity and high rates of missing data constrained our choices. We believe that repayment status and percent OPB remaining are sufficient to capture repayment pathways and likely represent the types of data most commonly available.

Finally, with respect to the observational sequences, note that the model is relatively naive—demographic variations and academic variations (e.g., institutional characteristics and degree types) are treated as exogenous and purposefully excluded from the model. This is due in part to the 2015 Federal Student Aid Supplement being an administrative match that precludes the collection of any time-varying covariates that are not in NSLDS after 2004. That said, many of these attributes are, by nature, relatively stable—either overall or from quarter to quarter—and therefore provide little utility from direct inclusion in the model.

HMM formalization

As noted above, HMMs allow a researcher to use an observed state sequence (in our case, discrete rather than continuous, and consisting of two channels of observed sequences) $y = (y_1, \dots, y_T)$ with observed states $m \in \{1, \dots, M\}$ to infer a hidden, or latent, state sequence $z =$

$(z_1, \dots, z_T)$ with states $s \in \{1, \dots, S\}$. Three matrices, which must be estimated, comprise the HMM: (1) the *initial probability vector*, (2) the *transition matrix*, and (3) the *emission matrix*.

In the *initial probability vector* $\pi = \{\pi_s\}$ of S length, π_s represents the probability of starting from the hidden state s

$$\pi_s = P(z_1 = s); s \in \{1, \dots, S\} \quad (1)$$

In the *transition matrix* $A = \{a_{sr}\}$, an S-dimensional square matrix, a_{sr} describes the probability of transitioning from the hidden state s at time $(t-1)$ to the hidden state r at time t; that is

$$a_{sr} = P(z_t = r \mid z_{t-1} = s); s, r \in \{1, \dots, S\} \quad (2)$$

In our HMM, we use a homogeneous model in which the transition probabilities a_{sr} are fixed over time.

In the $S \times M$ *emission matrix* $B = \{b_s(m)\}$, $b_s(m)$ is the probability of the hidden state s emitting the observed state m

$$b_s(m) = P(y_t = m \mid z_t = s); s \in \{1, \dots, S\}, m \in \{1, \dots, M\} \quad (3)$$

As mentioned above, within the HMM, certain states are modeled as latent and therefore unobserved (hidden). Hidden states are embedded within a first-order Markov process, such that movement between the prior hidden state and the next state is dependent only on the single, previous state $(t-1)$. Formally, this may be expressed by

$$P(z_t \mid z_{t-1}, \dots, z_1) = P(z_t \mid z_{t-1}) \quad (4)$$

Additionally, an observation at time t is dependent only on the current hidden state, rather than previous observations or hidden states:

$$P(y_t \mid y_{t-1}, \dots, y_1;\ z_t, z_{t-1} \dots, z_1) = P(y_t \mid z_t) \quad (5)$$

Multisequence and multichannel HMM

Equations 1 through 5 above describe a single-sequence single-channel HMM; extension of these to represent multiple sequences (i.e., separate sequences for multiple students) is, however, relatively straightforward. Instead of the single observed sequence y, we allow for N

sequences, represented as $Y = (y_1, \dots, y_N)^{\intercal}$ in which the observations $y_i = (y_{i1}, \dots, y_{iT})$ of each student borrower i take values in the observed state space. Importantly, observations are assumed to be generated by the *same* model, but each student has his or her own hidden state sequence. For multichannel sequences, there are C parallel sequences (in our case two such sequences, as noted above) for each student i. In this case, observations take the form $y_{itc}, i \in \{1, \dots, N\};\ t \in \{1, \dots, T\};\ c \in \{1, \dots, C\}$. Each channel is assumed to share a single common transition matrix A, but several (C-many) emission matrices, $B_1, \dots, B_C$, representing one for each of the channels.

Initial parameters

Before the HMM can be fit, the researcher must define the number of hidden states and initial values for the emission, transition, and initial probability matrices. Models of several hidden-state lengths were explored. To maximize both model fit and interpretability, a five-state model was selected. We encoded the three matrices with information describing some probable behavior of our model to improve the likelihood of convergence and model fit. Note that these matrices still had to be estimated but were simply initialized with likely starting values.

Most notably, for the transition matrix A, we initialized the matrix to a pseudo-Bakis model in which states were arranged in a roughly ordered sequence (e.g., it was more probable that a student would transition from state 2 to 3 than from state 3 to 2, but the transition from state 3 to 2 could be nonzero). In a true Bakis model, states move sequentially such that state 1 precedes 2, with a transition probability $p = 0$ for movement from state 2 to 1—the transition matrix for a true Bakis model is therefore upper triangular in form. Bakis and pseudo-Bakis models provide utility when the sequence of hidden states is assumed to show a temporal progression. The pseudo-Bakis model is initialized to

$$\mathrm{A} = \begin{bmatrix} .75 & .10 & .05 & .05 & .05 \\ .05 & .75 & .10 & .05 & .05 \\ .05 & .05 & .75 & .10 & .05 \\ .05 & .05 & .05 & .75 & .10 \\ .05 & .05 & .05 & .10 & .75 \end{bmatrix}$$

We initialized the model with several pseudo-Bakis matrices before selecting the transition matrix above, under which the Bayesian information criterion (BIC), a measure of model efficiency with respect to predicting the given data, was minimized (BIC = 522,742.8).

For emission matrix B, we let emissions be drawn from the following quarter periods: state 1, quarters 1-2; state 2, quarters 3-10; state 3, quarters 11-20, state 4, quarters 21-30; state 5,

quarters 31-50. In this way, we allow states to be ordered in an approximate temporal progression. As with the transition matrix, we tested several variations of the initial emission matrix before the above specifications were selected to minimize the BIC.

Parameter estimation

To estimate the unknown transition, emission, and initial probabilities, we use maximum likelihood estimation. The log-likelihood of the parameters $\lambda = \{\pi, A, \dots, B_C\}$ takes the form

$$\log L = \sum_{i=1}^{N} \log P(Y_i \mid \lambda) \tag{6}$$

in which Y_i represents the observed sequences in channels $1, \dots, C$ for student borrower i. Thus, for student i, the probability of the observed sequence conditioned on λ is given by

$$P(Y_i \mid \lambda) = \sum_{\text{all } z} P(z \mid \lambda) \cdot P(Y_i \mid z, \lambda) \tag{7a}$$

$$= \sum_{\text{all } z} P(z_1 \mid \lambda) \cdot P(y_{i1} \mid z_1, \lambda) \cdot \prod_{t=2}^{T} P(z_t \mid z_{t-1}, \lambda) \cdot P(y_{it} \mid z_t, \lambda) \tag{7b}$$

in which the hidden state sequence $z = (z_1, \dots, z_T)$ takes all possible value combinations of the hidden state space ($\{1, \dots, S\}$), and y_{it} represent the observed statuses of borrower i at t in channels 1 through C.

Inference

To make inferences using this model and observed sequences, we use the forward probabilities $\alpha_{it}(s)$, as specified in Rabiner (1989), which are the joint probability of hidden state s at time t as well as the observed sequences $y_{i1}, \dots, y_{it}$ (i.e., all the previous sequences up to the current sequence), given the model parameters λ. Backward probabilities $\beta_{it}(s)$ the joint probability of hidden state s at time t and sequences $y_{i(t+1)}, \dots, y_{iT}$, conditioned on λ. Using the forward and backward probabilities, the posterior probabilities of states can be estimated using the Baum-Welch forward-backward algorithm, which gives the probability of borrower i being in each hidden state for each time t, given the observed statuses of borrower i.[15] These are given by

$$P(z_{it} = s \mid Y_i, \lambda) = \frac{\alpha_{it}(s)\beta_{it}(s)}{P(Y_i \mid \lambda)} \tag{8a}$$

$$= \frac{\alpha_{it}(s)\beta_{it}(s)}{\sum_{s=1}^{S} \alpha_{it}(s)\beta_{it}(s)} \tag{8b}$$

From equations 8a and 8b, posterior probabilities can then be used to find the most probable hidden state at each time (as a local maxima).[16]

External review

The report benefited from the insights and expertise of Will Doyle of Vanderbilt University. Although he reviewed the report's methodology, neither he nor his organization necessarily endorse its conclusions.

Endnotes

[1] Statistics calculated from the National Postsecondary Student Aid Study of 2016 (NPSAS:16).
[2] Statistics calculated from the Beginning Postsecondary Students Longitudinal Study of 2004 (BPS:04/09). This analysis does not examine forbearances due to data limitations.
[3] Here, "experiencing distress" refers to experiencing state 3 as described in the "About the Data" section. In state 3, most borrowers are in distress—for example, in default or economic hardship deferments. However, some of the borrowers experiencing state 3 are in education deferments. These borrowers had other repayment characteristics that might indicate distress.
[4] For more details, see https://nces.ed.gov/datalab/powerstats/pdf/npsas2016ug_varname.pdf.
[5] For more details, see https://studentaid.ed.gov/sa/repay-loans/understand#grace-period.
[6] For more details, see https://studentaid.ed.gov/sa/repay-loans/understand/plans/income-driven.
[7] For more details, see https://studentaid.ed.gov/sa/repay-loans/default#default.
[8] Throughout the paper, "distress" refers to state 3.
[9] Borrowers could have been negatively amortizing for a host of reasons including, but not limited to, being in a repayment plan in which their monthly payments (or making monthly payments that) do not keep up with the interest that accrues on their loans, using an economic hardship deferment, or being delinquent on their loans.
[10] Students are considered independent for financial aid purposes if they are 24 years of age or older, are married, have children, or are veterans of the military.
[11] For more details on these data, see Nichole D. Smith and Michael A. Duprey, "2015 Federal Student Aid Supplement for the 1996 and 2004 Beginning Postsecondary Students Longitudinal Study Cohorts Data File Documentation (NCES 2018-409)," National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education (2017), https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018409.
[12] Parent PLUS loans were excluded from the analysis since the students themselves are not legally responsible for repaying these loans.
[13] For federal student loans, maturity date refers to the date a loan enters repayment, and once in the past, does not change as a result of subsequent administrative action like deferment, forbearance, or default.
[14] We chose quarters as the time unit because, since 2005, this is the frequency at which loan servicers are required to report or certify outstanding principal loan balances to NSLDS.
[15] Leonard E. Baum et al., "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The Annals of Mathematical Statistics* 41, no. 1 (1970): 164-71, https://projecteuclid.org/euclid.aoms/1177697196; Satu Helske and Jouni Helske, "Mixture Hidden Markov Models for Sequence Data: The seqHMM Package in R," *Journal of Statistical Software* (submitted 2017), https://arxiv.org/pdf/1704.00543.
[16] For a more detailed treatment of the procedures involved in inference and parameter estimation, see texts such as Iain L. MacDonald and Walter Zucchini, *Hidden Markov and Other Models for Discrete-Valued Time Series* (vol. 110) (Boca Raton, Florida: CRC Press, 1997); Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE* 77, no. 2 (1989): 257-86, https://ieeexplore.ieee.org/document/18626. For a more detailed discussion of the statistical package used for modeling, see documentation for the seqHMM R package, especially Helske and Helske, "Mixture Hidden Markov Models."