Res High Educ (2017) 58:449–467
DOI 10.1007/s11162-016-9430-2

The College Completion Puzzle: A Hidden Markov Model Approach

Dirk Witteveen[1] · Paul Attewell[1]

Received: 24 November 2015 / Published online: 9 August 2016

Abstract Higher education in America is characterized by widespread access to college but low rates of completion, especially among undergraduates at less selective institutions. We analyze longitudinal transcript data to examine processes leading to graduation, using Hidden Markov modeling. We identify several latent states that are associated with patterns of course taking, and show that a trained Hidden Markov model can predict graduation or nongraduation based on only a few semesters of transcript data. We compare this approach to more conventional methods and conclude that certain college-specific processes, associated with graduation, should be analyzed in addition to socio-economic factors. The results from the Hidden Markov trajectories indicate that both graduating and nongraduating students take the more difficult mathematical and technical courses at an equal rate. However, undergraduates who complete their bachelor's degree within 6 years are more likely to alternate between these semesters with a heavy course load and the less course-intense semesters. The course-taking patterns found among college students also indicate that nongraduates withdraw more often from coursework than average, yet when graduates withdraw, they tend do so in exactly those semesters of the college career in which more difficult courses are taken. These findings, as well as the sequence methodology itself, emphasize the importance of careful course selection and counseling early on in student's college career.

Keywords College completion · COURSE-TAKING · Academic momentum · Quantitative methodology · Longitudinal analysis

✉ Dirk Witteveen
dwitteveen@gradcenter.cuny.edu

[1] The Graduate Center, The City University of New York, 365 Fifth Avenue, Room 6112.04, New York, NY 10016, USA

Introduction

This study builds upon the availability of coded undergraduate transcript data and the recent development of computationally intensive methods for analyzing sequences and trajectories over time in order to address the college completion puzzle. We use Hidden Markov Modeling to analyze longitudinal (transcript) data in order to predict degree completion and noncompletion. The main function of this modeling technique is to identify the so-called 'hidden states' that are associated with both static observable states and—harder to observe—trajectories leading to particular outcome states. A Hidden Markov Model (HMM hereafter) becomes particularly useful for sociologists if one assumes that the complex—yet observable—sequences have underlying dependence associated with the outcome variable, such as in college completion patterns.

Various versions of HMM's have been introduced by Langeheine and Van der Pol (2002), Elzinga et al. (2007), and Vermunt et al. (2008) in the methodological literature. Furthermore, HMM's have been applied to a relatively small sample (N = 550) of violence profiles to reveal trajectories of batterers by Ip et al. (2010). Although these studies are valuable technical introductions to HMM's, their focus remains on sequences *without* structurally linking these to a dependent (outcome) variable.

We present a substantial extension of the existing approach—a Hidden Markov Model to reveal states and trajectories among a sample of U.S. 4-year college students, and we also use the HMM to predict outcome states (graduation or nongraduation)—a technique that is much more suitable for social science research. We contrast both our techniques and findings with more conventional modeling approaches to understanding trajectories toward college graduation. We show how identified states and trajectories can predict graduation patterns *beyond* frequently used socio-economic, demographic, and precollege background information.

The College Completion Puzzle

The higher education system in America is characterized by widespread access to college but a low rate of degree completion. After graduation, many high school students proceed immediately to college (68 % in recent cohorts), but a substantial proportion of them do not complete their degree programs after they enter: about 63 % of entrants to 4-year colleges complete a bachelor's degree in 6 years (Aud et al. 2013; Radford et al. 2011). Although other OECD countries display similar graduation rates, such as Finland (58 %) for its university sector, Sweden (54 %), France (64 %), and Norway (67 %), as reported by Adelman (2009), the US college dropout rate is lower than many policy makers would desire and therefore deserves attention.

A large body of research has examined the determinants of student degree completion. In broad terms, one can identify four approaches. The first emphasizes the importance of students' background characteristics such as family income or wealth, parents' educational attainment, plus students' high school coursework, and their scores on skills tests—factors that pre-date entry into college but are associated with the likelihood of college graduation (Achieve 2004; Chen 2005; Complete College America 2011; Horn and Kojaku 2001).

A second stream of research examines nonacademic circumstances that undergraduates cope with after entry to college: the amount of financial aid (Perna 2010; Perna and Li 2006; Schuh 2005; St. John et al. 2000; Wyner et al. 2007); the number of hours of paid work they undertake (Bozick 2007); whether they have family responsibilities (Bean and

Metzner 1985); whether they live on campus or commute (Bozick 2007)—and links these to degree completion. These factors are conceptualized as competing obligations or ongoing stresses that affect an undergraduate's ability to finish a course of study.

A third research approach focuses on differences in graduation rates across institutions (Hess et al. 2009; Schneider and Yin 2012). Some take these differences to imply that certain colleges are more efficient than others at graduating students, leading these scholars to argue that students should try to attend the most selective college possible, since this will enhance their chances of graduating (Bowen et al. 2009).

The Hidden Markov approach presented below belongs to a fourth body of research that seeks answers to the noncompletion puzzle in undergraduates' early semesters in college. Its core idea is that some have experiences early on in college that discourage them and lead to stopping out or to dropping out, while others experience successes that build motivation and commitment and make it more likely that the student will persist and complete. Tinto (1993) pioneered this approach, conceptualizing the issue as a matter of "fit" between each student's aptitudes and the demands made by the college. A later wave of research conceptualized early experiences of undergraduates in terms of their *academic momentum* rather than students' integration. Academic momentum is represented by the number of course credits and GPA that an undergraduate accumulates during the first year at college, and their trajectory over time. Adelman (1999, 2006) demonstrated that early momentum predicts degree completion, using transcript data from nationally representative panel studies of college students, above and beyond students' academic preparation and socio-demographic background. Subsequent research has largely confirmed Adelman's insights (Author date; Complete College America 2011).

When understanding college completion (or dropout) patterns, particular courses and their grades also play an important role. With data from the US Department of Education (2004), the proportion of withdrawals in college grew substantially between 1970s and through the 1990s (Adelman 2004). This study also showed that both withdrawals and failures are concentrated in math courses (mainly algebra) and in remedial courses (writing and reading). Other researchers have been concerned with course-taking profiles and cumulating patterns of withdrawals and failures, but for high schools (i.e. Heck et al. 2004).

The availability of coded transcript data that detail student progress toward the degree (specific courses taken, grades obtained, course-loads each semester, periods of stopping out, and transfers between colleges) enables further research that seeks to understand fine details of undergraduate academic trajectories using new data-mining techniques such as Hidden Markov models. Certain courses of study may be more difficult than others; particular types of courses may act as gatekeepers and have higher failure rates than others. From this viewpoint, student progress toward a degree may be thought of as a kind of track-and-field race over hurdles, where failing to clear certain academic hurdles or encountering difficult course sequences or getting poor grades in particular courses lead some students to redirect their efforts, sometimes into another major and sometimes causing them to stop out of college or to transfer.

Data

Our analyses draw upon a study known as the 'Beginning Postsecondary Students Longitudinal Study' (BPS), directed by the National Center for Education Statistics (NCES). It tracks a nationally representative cohort of first-time freshmen for 6 years after their initial entry into college in 2004. Each student reported every college attended during this period; both dual enrollment and transfers from one college to another (Wine et al. 2011).

BPS researchers subsequently obtained and coded transcripts from each college a student attended, creating a dataset known as the “2004/2009 Beginning Postsecondary Students Longitudinal Study Restricted-Use Transcript Data Files” or PETS data (NCES 2011). Our analyses use these transcript data to measure student progress semester by semester, including transfers and final degree attainment. Our conventional analyses—the (multinomial) logistic regressions—use weights developed by the BPS staff to adjust for panel attrition and nonresponse. Panel weights are however not applied in our Hidden Markov models.[1] Yet in a separate series of conventional logistic regression analyses (in which we are interested in background factors such as race and gender), the BPS panel weights are applied to adjust for sampling errors. In addition, BPS replaced all missing data using multiple hot-deck imputation.

Within the BPS transcript sample, 8980 students (a rounded figure) registered in a 4-year college for their first semester in Fall 2004.[2] These first-time freshmen at baccalaureate-granting institutions form the sample for our analyses. Using these longitudinal data, we address the question: How can students who graduate within 6 years be distinguished from those who did not graduate within the same time frame?

Independent Variables: Semester Characteristics

Each data point consists of a student by semester observation, measured as a set of values of selected variables in a particular fall or spring semester for all 4 years of observation. All the BPS variables used here are discrete (nominal or ordinal) or they are reconstructed as such. Hence, every combination of variable values ($m_{1\ldots n}$) creates our sequence of ‘semester observations’ ($X_{1\ldots n}$). Table 1 lists all predictors as well as their averages across all 8 semesters for graduates and nongraduates separately.

As seen in Table 1, nongraduates are more likely to stop out and to enroll less frequently. They also fall behind graduates in terms of credits earned, average GPAs, and are more likely to take remedial coursework and to withdraw from or fail courses. Surprisingly, they also on an average have a higher degree of involvement in science, technology, and math (STEM) courses than graduates. These dimensions form the basis of the stochastic models we will build to predict graduation.

Methodology

The terminology of Hidden Markov Modeling, as introduced in this paper, deviates from conventional techniques such as logistic regression. Here, we will introduce some central terms in HMM, before providing a more detailed description of this method in the following sections.

[1] Panel weights are not conventionally used in the construction of the HMM itself (the hidden states). An HMM looks at variation over time within individuals’ sequences rather than representativeness of samples to a larger population.

[2] Around 8190 of these sampled students were 18 or 19 years old when they entered a four-year college for the first time (790 students were 20 years or older). At the urging of one reviewer, we reran the Hidden Markov model omitting students who were 20 years or older. These reworked analyses yielded similar results in terms of state description, transitions probabilities, and prediction accuracy, and are available upon request.

Table 1 Predictor variables from 4-year college transcripts

Independent variable	Averages across 8 semesters		Scale on semester level
	Nongraduates	Graduates	
Student enrolled this semester	68.7 %	97.1 %	Dichotomous
Number of credits attempted[a]	12.5	14.5	5 Categories
Cumulative credits attempted[b]	48.1	67.6	27 Categories
Number of credits earned	9.5	13.9	5 Categories
Weighted GPA by semester[c]	2.24	3.15	4 Categories
Cumulative GPA	2.33	3.13	4 Categories
Student took remedial English	4.4 %	1.2 %	Dichotomous
Student took remedial Math	14.7 %	4.6 %	Dichotomous
Student took remedial 'other'	33.3 %	20.4 %	Dichotomous
Student withdrew from 1 or more courses	23.1 %	8.9 %	Dichotomous
Student failed 1 or more courses	43.1 %	5.9 %	Dichotomous
Student took at least 1 STEM course[d]	64.4 %	58.1 %	Dichotomous

[a] The 'number of credits' variable was standardized across colleges by NCES

[b] Z-score distribution was used to create effective bins

[c] A standardized GPA (per credit) was used to create this variable. The following NCES codebook was used to decide on these criteria: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2012162rev

[d] STEM refers to courses in Science, Technology, Engineering, and Math. However, remedial or developmental courses were not counted as STEM courses

HMM seeks to identify processes unfolding over time. At any moment in time, each individual in the analysis is said to reside in one of the small number of 'states,' and any individual may move from one state at one period into a different state at the next or may remain in the same state. So, individuals move through a sequence of states over time. Each state is a kind of latent variable constructed by the HMM out of measured variables. For example, one state might be characterized as a 'STEM-withdrawing state'; another might be a 'low course-taking state.' What is of interest to the analyst is first, identifying the states that best describe the process over time using an 'observation matrix'; second, understanding transition probabilities: the likelihood of moving from one state to another at any point in time; third, fitting longitudinal data to particular sequences of states; and fourth, associating particular sequences of states to overall outcomes, such as graduating vs. not graduating.

Hidden Markov Models: An Overview

A Markov Model is a probabilistic model that predicts the trajectory of certain variables over time. Collectively, those variables are usually termed "the state of the system." One important assumption, known as the *Markov Property*, is that the state of the system at time t depends upon, or is only affected by, the state of the system at the immediately prior point in time—time $(t - 1)$—and not by values earlier in time such as $(t - 2)$ or $(t - 3)$. In other words, any effects of earlier states of the system are conceptualized as incorporated into the state of the system $(t - 1)$ immediately prior to the current state of the system.

In a simple Markov model, the state of the system might be represented by one or more observed variables. So, for example, the state might be whether an undergraduate is enrolled or not enrolled in college, semester by semester, as a probabilistic process whereby the measured outcome varies over time and is affected by its value at the prior semester. That kind of simple Markov model might therefore summarize the enrollment trajectories of students.

However, a *Hidden* Markov model (HMM), which is used here, involves a more complex situation where the state is *not* directly observed. Instead, the state of the system at any point in time is conceptualized in terms of an ensemble of several "hidden states"—analogous to latent variables. These hidden states are linked in a Markov process, such that the probability of a particular hidden state at time t is a function of the probabilities of a set of states at time (t − 1). Parameters known as transition probabilities describe the likelihood of individuals moving from one state to each other state in the next time period. In HMM modeling, the hypothesized values of the hidden states (represented as Z in the formulae below) are used to predict the probability of observed variables $X_{0\ldots n}$ rather than vice versa, as presented visually in "Appendix 1."

The HMM algorithm involves a form of 'machine learning': through an iterative process, a program estimates certain parameters in the model that yield predictions of the values of observed variables. Over many cycles, the algorithm converges to represent the best fit between the model and the data. This 'training of the model' produces a stochastic structure that explains change over time (in this case, semester by semester) and links system states to observed variables such as grades, number, and types of courses taken.

The core of the HMM model is the following formula that predicts the probability of finding a particular *sequence* (of length n) as follows:

$$P(X_{0\ldots n}) = P(Z_0)\prod_{t=1}^{t-1} P(Z_t|Z_{t-1})\prod_{t=0}^{t-1} P(X_t|Z_t) \tag{1}$$

P($X_{0\ldots n}$) is the probability of an observation in a Hidden Markov sequence and is equal to the probability of any initial hidden state (Z_0), multiplied by the product of a temporal chain of conditioned hidden states $(Z_t|Z_{t-1})$—the *transition probability*—and by the product of observations in that same time frame $(X_t|Z_t)$—the *emission probability*. Hence, the algorithm predicts an *observation sequence* using the sequence estimates of a current *state* and the immediately prior *state* (Eq. 1). The emission probability—$P(X_t|Z_t)$—indicates the conditional distribution of observations from a specific state.

Knowing this, one could predict a current and future *observation* or *hidden state* using the estimation of our past observation and state. Calculating these probabilities requires three functions that can be recognized in Eq. (1): an initial state probability P(Z_0), a state transition function P$(Z_t|Z_{t-1})$, and an observation function P$((X_t|Z_t))$.

Using those three matrices, one can also calculate the probability of finding future (hidden) states (up to Z_n), where n is the length of our observation sequence $(X_{0\ldots n})$, by a so-called *forward–backward algorithm* (based on Baum et al. 1970):

$$P(Z_{t+1} = z|z_{1\ldots t}, x_{1\ldots t}, x_{t+1}) = \sum_x P(Z_{t+1} = z|X_{t+1} = x)P(Z_{t+1}|z_{1\ldots t}) \tag{2}$$

The noncapitalized x's and z's refer to *past* observations and hidden states. The bold z's indicate specific values of z (a hidden state).

Any future state—Z_{t+1} in Eq. 2—is a function of the actual observation (x_{t+1}) at that time, as well as previous series of both the observations and hidden states $(x_{1\ldots t}, z_{1\ldots t})$. The right side of Eq. (2) therefore indicates the summation of the probability of finding a particular hidden state in the *next* time period ($Z_{t+1} = z$), given the actual observation ($X_{t+1} = x$), multiplied by the transition function: $P(Z_{t+1}|z_{1 \ldots t})$.

Deciding on the appropriate number of states is one early task in any HMM analysis. Typically, a researcher tries a model with two, three, four, or more states and decides on the best number according to the goodness-of-fit of these alternatives. This method of optimal state selection seems most appropriate to the social sciences, in which one aims for realistic—meaningful—labels that can describe a hidden state. Even though latent states are represented as probability structures across several predictors, they should preferably be distinct, recognizable, and informative. If adding or dropping a state reduces these conditions, one has reached an optimum number of states for the model. Alternatively, one can use a Bayesian Information Criterion (BIC) to choose the number of states in the model (Scott 2002; McLachlan and Peel 2000). In practice, however, even when using a BIC, one would still evaluate models, using different numbers of states, and inspect the qualitative aspects of the model (e.g., Ip et al. 2010). Ideally, both approaches should lead to the same decision.

As mentioned earlier, states are not directly observed; they are hypothesized latent variables. However, one *can* estimate a state. Hidden states are complex functions of multiple measured variables (the latter known as 'features'). This means that, in theory, a complete semester observation (X_t) can take any form; a series of discrete or finite numbers. The algorithm of an HMM estimates the probability of observing certain combinations of feature values at ($X_{0 \ldots n}$), while simultaneously estimating the latent variable $(Z_{0 \ldots n})$ in the same time frame. Therefore, the *hidden* states $(Z_{0 \ldots n})$ of the Hidden Markov Chain can be described by the probabilities of our directly observed chain of discrete and nondiscrete feature values—e.g., GPA, number of credits attempted. Consequently, after an HMM has been fitted successfully to data, one can observe what features are associated with each hidden state, to gain a retrospective sense of the meaning of each state. But rather like components obtained from principal components or factor analysis, a conceptually confusing mix of measured variables or features may contribute to each state. For a general introduction to HMM see Rabiner (1989) or Stamp (2015).

HMM Assessment and Predicting Outcomes

Building an HMM takes several steps: variable selection, state description evaluation, and estimating the number of states from the key activities. This 'analysis phase' leads to a trained model that consists of the initial state distribution (probabilities of being in one of the states in the first or *initial* semester; Z_0), the transition matrix (a probability matrix with the likelihood of being in a certain state Z_{t+1} given Z_t), and an observation matrix (probabilities of finding certain combinations of observed feature values in each latent state). In this phase, we attempt to detect and describe educational trajectories.

Our analysis takes an additional step, *after* HMM sequences are optimized, which makes it possible to predict a dichotomous, sociologically relevant, dependent variable. The logic of using an HMM to predict an outcome of interest, such as graduation from college, contrasts with both more conventional modeling approaches such as logistic regression and with previous applications of HMM's using social science data.

An HMM does *not* directly model or predict a dependent or outcome variable. Instead, for a theoretically important dichotomous outcome such as completed a degree (yes/no), one first separates individuals in the sample according to their value on this outcome of interest. The researcher therefore creates two sequence samples that will be analyzed separately, each with its *own* Markov chain. In our case, one consists of students who finished their degree within the 5 years of the study, and the other contains of only noncompleters. We therefore create (train) two completely separate HMM's: one for the graduates and another for nongraduates.

Machine learning usually employs a form of replication known as cross-validation to assess the validity and accuracy of any model (comparable to the R-square). All cases in the data are randomly assigned into either a *training sample* or *test sample*. In our case, 70 % of the college graduates and nongraduates were randomly selected for training and the remaining 30 % were assigned to a test sample. This leads to two test samples (one for college graduates and another for nongraduates) which are held separate and play no role in developing these HMM's.

The test sample cases that were held back are then used to test the predictive accuracy of the two trained HMM models which were previously created using only the training data. For each individual that we withheld for the test sample, a log-likelihood of the sequence under *both* trained HMM's was produced. The highest log-likelihood between the two HMM's should be considered the "classification" on the dependent variable: the prediction whether that student is a graduate or a noncompleter. The comparison of likelihoods and log-likelihoods for classification purposes is a standard procedure in computer science and machine learning. For an introduction to using log-likelihood as a classification procedure, see Duda et al. (1973, 2000).

An important additional technique applied in our models is to classify students from the test sample with *shorter* sequences of observed semester validation data in the HMM's. A strong model should effectively recognize a graduating or nongraduating student after only 1 or 2 years of college transcript information.

In sum, all models presented are assessed by their state description efficacy, followed by testing the accuracy of prediction for graduating, nongraduating, and all students together.[3]

Analysis

Determining Predictors and Number of States

An HMM ideally uses a small number of predictors because a simple model improves the ability to assess and replicate the study. In addition, using a relatively small number of predictors increases the chance of effectively interpreting a latent state; fewer, yet unique, feature values capture the essence of 'being in a state' as a college student. Using a limited set of potentially predictive variables is common in HMM's in social science research. However, it should be noted that, in theory, the initial model estimated by an HMM *could* be chosen at random, with an unlimited number of predictors (Rabiner 1989).

In order to have an indication of the relative importance of our semester-based predictors, all 12 candidate variables were initially included in an HMM one-by-one (Table 1). We used the Bayesian Information Criterion (BIC) of each of these single-variable HMM's to rank the predictive power of the variables. The high importance of student enrollment in a given semester for both graduating and nongraduating is not surprising, since one has to attend in order to graduate. We therefore decided to eliminate this as a separate variable and instead to incorporate this into every other variable as a

[3] Murphy's (2002, 2005) *Matlab* toolbox is used for all HMM calculations. See Appendix B (for HMM training) and Appendix C (for classification).

Table 2 Observation probabilities by estimated state among Graduates and Nongraduates

Independent variables	Graduates						Nongraduates					
	State 1		State 2		State 3		State A		State B		State C	
	Prob.	Norm prob. (%)	Prob.	Norm prob. (%)	Prob.	Norm prob. (%)	Prob.	Norm prob. (%)	Prob.	Norm prob. (%)	Prob.	Norm prob. (%)
Failed a course	0.009	26.3	0.008	23.0	0.017	50.7	0.093	38.5	0.063	26.1	0.085	35.4
Withdrew a course	0.031	33.7	0.025	27.5	0.036	38.8	0.026	29.4	0.025	28.6	0.037	42.0
Took STEM	0.109	23.8	0.186	40.5	0.164	35.7	0.124	35.0	0.082	23.2	0.149	41.8
Math Remedial	0.009	33.9	0.006	23.8	0.012	42.3	0.011	26.0	0.013	28.8	0.020	45.2
Credits attempted												
<12	0.043	40.5	0.036	33.7	0.027	25.8	0.135	47.2	0.078	27.3	0.073	25.5
12–14.99	0.112	34.5	0.114	35.2	0.098	30.3	0.093	39.4	0.061	25.9	0.082	34.7
15–17.99	0.074	21.6	0.121	35.2	0.149	43.2	0.058	40.3	0.031	21.8	0.054	37.9
>18	0.061	43.5	0.045	32.5	0.034	24.0	0.014	38.0	0.007	19.6	0.016	42.4
Absent/Not enrolled	0.003	11.6	0.008	31.9	0.014	56.5	0.105	51.9	0.028	13.6	0.070	34.5
Other factors	0.024	40.6	0.018	30.1	0.017	29.3	0.045	44.1	0.028	27.0	0.030	28.9
State label	"High credits"		"High STEM"		"STEM/withdrawing"		"Low activity"		"Low STEM"		"STEM/high credits"	

The 'Prob.' distribution refers the *unstandardized* associations: the emission probabilities separate value: ‘not enrolled this semester.’ Taking remedial English is remarkably predictive in this particular stochastic model, even though only a small percentage of the students take this course. Another noteworthy finding is that more or less the same variables are predictive for both graduating and nongraduating students in terms of BIC-rank (and -2 log-likelihood rank).

Next, we examined various sets of variables in order to optimize the state descriptions in terms of their coherence. The trade-offs between estimating fewer or more states, on the one hand, and the process of dropping and adding predictors to the model, on the other hand, led us to estimate the HMM containing 6 of the 8 strongest variables listed in Table 1. Moreover, a “three-state solution” turned out to be most effective in creating coherent and distinct state descriptions.

Interpreting the Observation Matrix

The observation matrix is the connection between the observed world of college transcript variables and the logic of hypothesized latent states, whereby each predictor category equals a probability of being associated with one of the three estimated states. This matrix can be used to derive in which states students can possibly be during a semester. Thus, we arrive at state descriptions in terms of probability associations with each predictor value. Together, these characteristics form narrative “state labels” which are listed in the bottom row of Table 2.

Since we trained two distinct HMM’s, one for graduates and one for noncompleters, we analyze their two observation matrices separately. Organized by latent state, Table 2 indicates the emission probabilities; the unstandardized effect size of a certain feature on the state description (“Prob.”). This is a *joint distribution*, defined by the probability of observing a particular feature value, given the estimated underlying hidden state: $P(X, Z) = P(X|Z)P(Z)$. Furthermore, the *normalized probabilities* reflect the probability of observing a particular feature value, given the fact that we estimated our *three specific* latent states: $P(Z_i|X)/\sum P(Z|X)$. These probabilities are translated into percentages and add up to 100 % row-wise for both graduates and nongraduates.

Several contrasts can be seen between States 1 thru 3. First, STEM-taking is of great importance when describing the three states among graduating students. For instance, when observing a student who takes STEM, the normalized probability of that student being in State 1 is only 23.8 %, compared to a 40.5 % chance of being in State 2. Another important feature is the number of credits attempted in each semester. Taking more than 18 credits in a particular semester is associated with a 43.5 % chance of being in State 1, whereas the chances of being in the other two states are substantially lower: 32.5 % (State 2) and 24.0 % (State 3). Lastly, whereas State 2 can be described as a ‘regular’ state in which a student takes STEM and attempts a ‘normal’ number of credits, a student in State 3 seems to struggle a bit: the chances of Math remedial, stopping out, withdrawing a course, and above all, failing a course are higher in State 3 than in any other state.

Nongraduating students are described with exactly the same variables because this allows us to test both our models with new test data. Similar to graduates, STEM-taking and the number of credits attempted in a semester are important features that describe State A thru C. However, other factors contribute to the description of nongraduation states, such as failing or withdrawing a course and taking remedial Math (Table 2).

Failing a course is an important characteristic of all three states (around 35 %—39 %), but to a lesser extent of State B (26 %). Withdrawing takes place quite frequently across the three states, but observing this is most associated with State C—a 0.42 probability.

Furthermore, State A is clearly associated with a high probability of not being enrolled in that semester and, if enrolled, students in this state enroll in fewer credits; taking less than 12 credits (0.47 probability) and remedial Math (0.26 probability). Therefore, we labeled this state as "low activity." Lastly, State C can be identified as the "STEM-taking"-state; a 42 % chance of being in this state, as opposed to 35 % (State A) and 28.6 % (State B). An important difference between state B and C, however, is the number of credits attempted; a student who takes more than 18 credits, has a 19.6 % chance of being in state B, but a 42.4 % of being in state C. In other words, State C (among nongraduates) is quite comparable to State 1 (among graduates), with the important exception of being far more likely of taking many credits in one semester, instead of a high probability of withdrawing.

Examining College Trajectories: The Transition Matrix

After estimating the composition of hidden states, we can analyze probabilities of future states given our knowledge of the current state. In other words, if during any point in the college career a student displays a high likelihood of being in a particular state (1 thru 3 or A thru C), what is the chance of being in a certain state during the next semester? Again, it is important to keep in mind that we do not directly observe the hidden state, but instead, we rely on the *probability* a student 'being' in a latent state.

Table 3 indicates the initial state distribution—the probability of being in one of the three states across all observations—and the transition matrix for both the sequences of graduates and nongraduates. The latter is a 3 × 3 since we have defined three states in which students can possibly be in during a semester. We can model the probability of being in a certain state Z_t, given Z_{t-1} (see Eq. 2).

With regard to graduating students, the "high credits" State 1 is least likely to be a starting state. And, once in State 1, move quickly into a different state, such as State 3—the "STEM/withdrawing" state (79 % chance). Furthermore, once in States 2 or 3, the chance of moving back into a situation of taking high credits (State 1) is unlikely.

The crux of a "path to graduation" should be found in the combination of states 2 and 3. Starting in State 3 ("STEM/withdrawing") is most common, but once a student is in that

Table 3 Initial states and transition matrix for graduates and nongraduates

Initial states		Transition probabilities		
		State 1	State 2	State 3
Graduates				
State 1	0.077	0.142	0.071	0.787
State 2	0.306	0.172	0.336	0.492
State 3	0.617	0.161	0.566	0.273
Initial states		Transition probabilities		
		State A	State B	State C
Nongraduates				
State A	0.410	0.422	0.143	0.435
State B	0.412	0.445	0.438	0.117
State C	0.178	0.052	0.486	0.462

For graduating students*: state 1* high credits, *state 2* high STEM, *state 3* STEM/withdrawing

For nongraduating students*: state A* low activity, *state B* low STEM, *state C* STEM/high credits state, moving to State 2 (just “high STEM”) is more likely than staying in State 3. This indicates a ‘slow start’ with regard to STEM-taking. However, being in State 2 means that there is almost a 50 % chance of being (back) in State 3 during the next semester.

The fact that graduating students alternate between the estimated state 2 and 3 has two implications. First, graduating students rarely take STEM courses in combination with a large number of credits; they either withdraw from them, supplemented with remedial Math courses (State 3), or they take fewer credits overall (State 2). Moreover, the price that has to be paid in State 3 is a high chance of failing a course of about 50 % in terms of normalized probability. This suggests a trade-off for getting through the ‘harder’ courses in college (or a “strategy” when assuming that students choose their courses more or less rationally before the start of each semester). The high probabilities of moving between State 2 and 3 suggests that students alternate between the two ways of getting through a semester with a STEM course. Hence, the states and their transitions do not reveal an exclusive ‘paths to graduation,’ but rather substitute that can be considered to characterize every semester.

How are the academic activity paths of nongraduates developed? Among this group, a quite obvious conclusion can be drawn from one of the state descriptions: a high chance of not being enrolled is a clearly defined state (A) that is logically associated with non-graduation. However, once in this state, moving to a course-intense state the next semester is 57.8 %—the combined probability of being in state B or C. Once there, moving back and forth between being enrolled and not enrolled is common when entering State B (45 % chance), but very unusual when entering State C (5 % chance). This pattern between State A and B may indicate a “stop-out pattern”: many students who were not enrolled in the previous semester do in fact return in the next semester, after which students are then not enrolled in the following semester. Importantly, the ‘stickiness’ of state C among non-graduating students—the state that looks most similar to a state found among graduates (“High STEM”)—is remarkable and not observed among graduates. This suggests that students who graduate within 6 years are better able to *maneuver* through the more difficult courses and tougher semesters.

Predictive Power

Accuracy for Different Spans of Time

After describing the states and the transition trajectories of graduating and nongraduating students, we can now assess the predictive accuracy of each HMM obtained from training data using the held back ‘test data’ (30 %). Given a sequence of semesters with specific features, what is the probability of that chain being associated with graduating within 6 years or not? For answering this question we use the language of machine learning, in the sense that we phrase the results from our model in terms of ‘predictions’ rather than ‘association,’ as is common in inferential statistics.

Figure 1 below displays the accuracy rates of our best model as presented in this paper. The accuracy rates are organized by the number of semesters that were used to test the model (up to 8 semesters). Hence, longer sequences of semesters lead to fewer classification errors. This explains the general positive slopes.

The diamond line indicates the *overall accuracy* of the HMM, which is the average of correctly predicted graduating *and* nongraduating students in the test sample. The prediction accuracy is the total percentage of students correctly classified in the test data. This

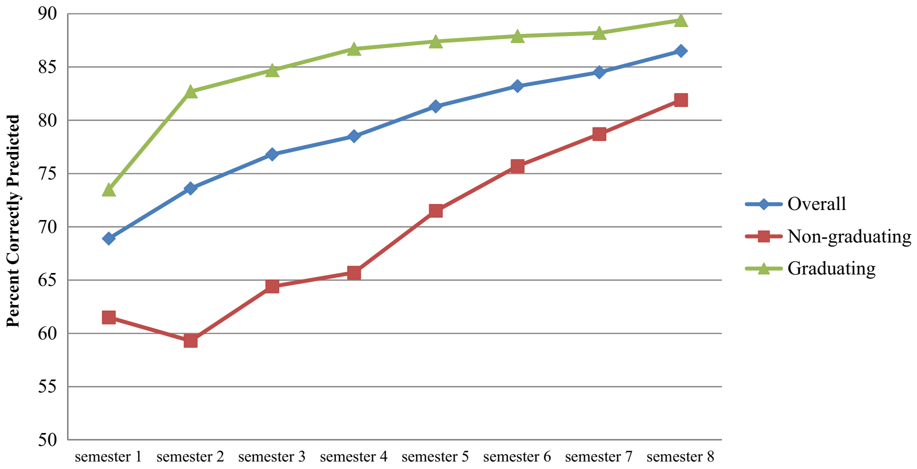


Fig. 1 Accuracy rates of HMM's for longer and shorter sequences

varies from 87 % when providing the model with transcripts for 8 semesters, down to 69 % when providing the model only with the first semester's transcript.

Two conclusions may be drawn from the discrepancy in correctly predicting graduates (triangle) version nongraduates (squared). First, a large group of students surprises us. They seem to fall behind in the first couple of semesters, and their transcripts resemble those of drop-outs, but they catch up later on—roughly speaking around the fifth semester. Presumably, these students switch to a state that is associated with a path to graduation. Second, the state descriptions already indicated that a student who is likely to be in State 3 ("STEM/withdrawing"), among graduates, could in principle also be in State C ("STEM/ high credits"), among nongraduates, because the only qualitative difference is the lower chance of withdrawing from a course. This helps to explain why our HMM has difficulties distinguishing some of these student-states, leading to somewhat lower accuracy rates for nongraduates.

Hidden Markov Approaches and Logistic Regressions

The best performing HMM in this study describes a process of college decisions and experiences; *proximal causes* of college completion and dropout. While ignoring any knowledge about a student's previous and current socio-demographic condition that traditional sociological studies have long used to predict college graduation (e.g., parent's education, number of dependent children), the Hidden Markov model reveals a separate sociological process: the college trajectory effect. This does not, however, imply that graduation patterns found using an HMM are necessarily independent of socio-demographic effects; this is an empirical question.

To examine whether the states of an HMM model do reflect sociological variables, we ran a multinomial logistic regression on nongraduating students for each of the 8 semesters observed, with the three states as outcome variables, and the "inactive/low activity" (State A) as the reference category. We used a set of commonly used socio-demographic predictors in order to reveal associations between socio-economic factors and college states:

gender, race, parent's education, English as first language, high school GPA, having dependent children, hours worked per week, and age at start of college. The results (not shown, but available from the authors) only indicate a consistent significant effect of gender, predicting male students being more likely to be in the "STEM/high credits" state. Other demographic and high school variables were not significant predictors of the HMM states. We interpret this to mean that the states in our HMM do not *just* represent socio-demographic background; the graduation trajectory as found here is a process that is largely independent of a student's demographics.

Not only does an HMM add semester-specific variables and semester-states to the college graduation puzzle, it also produces a high accuracy of future predictions that can compete with a logistic regression approach. Table 4 presents the confusion matrices of a logistic regression on college graduation, using the *same sample* and set of commonly used socio-economic predictors, as well as a series of Hidden Markov Models with different numbers of semester transcripts used. The accuracy rates of the HMM were calculated using only 30 % of the original sample.

As seen in Table 4, after using just two semester transcripts (1 year of college), the HMM predicts almost 74 % of these test cases correctly, with a 59 % accuracy for non-graduates. Moreover, when adding one college year of transcripts at a time to our model input data, the overall accuracy increases to 79, 83, and 87 % (four college years). In contrast, the accuracy rate of the more conventional logistic model predicts about 70 % of cases correctly. Only 43 % of the 'less frequent case' (noncompletion) was correctly classified by a logistic model. In sum, an HMM performs very well on predicting the rarer

Table 4 Confusion matrices of a logistic regression and different versions of HMMs on graduation within 6 years

	Prediction			Classification %
	False	Correct	Total	
Logistic regression				
Did not graduate	1813	1672	3485	48.0
Did graduate	839	4657	5496	84.7
Total	2652	6329	8981	70.5
HMM—1 year transcript data				
Did not graduate	425	620	1045	59.3
Did graduate	285	1364	1649	82.7
Total	710	1984	2694	73.7
HMM—2 year transcript data				
Did not graduate	358	687	1045	65.7
Did graduate	220	1429	1649	86.7
Total	578	2116	2694	78.5
HMM—3 year transcript data				
Did not graduate	254	791	1045	75.7
Did graduate	199	1450	1649	87.9
Total	453	2241	2694	83.2
HMM—4 year transcript data				
Did not graduate	189	856	1045	81.9
Did graduate	175	1474	1649	89.4
Total	364	2330	2694	86.5

Logistic regression covariates: gender (male), race (black, Latino), parental education (high), high school GPA, number of children, work hours, age, and high school math (trigonometry, algebra I/II, calculus, with reference precalculus) displayed significant association with the dependent variable in the expected direction. Panel weights were applied. Summary statistics: N = 8981, $R^2 = 0.143$

case—nongraduation—in comparison to the logistic regression. Without using any of the socio-demographic factors of college students, HMM predictions display much higher specificity rates than the logistic regression.

The implication is that both a logistic regression, containing socio-demographic independent variables, and a trained HMM, with just the first two semesters as data, predict college student's graduation with an accuracy of around 70 %. These comparisons of confusion matrices apply exclusively to the predictive power of logistic regressions and HMMs, indicating the competitiveness of the latter with regard to model fitness (not its superiority). Since our observed states are not associated with demographics (except gender), yet still deliver satisfactory prediction accuracy, we argue that the HMM provides the educational researcher with an additional technique to capture the logic of college experiences. In other words, modeling the college trajectory itself—through an HMM—provides a useful sociological tool to understand trajectory factors of college graduation beyond previously found structural or sociological factors. It enables the researcher to explain a phenomenon such as college completion with more fine-grained understanding of the course and grade sequences upon which these outcomes are based.

Discussion

Scholars have long sought answers to the question of why so many American undergraduates fail to complete their degrees. For this puzzle, we took an approach that is rarely applied in the social sciences: Hidden Markov modeling. Based on the empirical results, our claim is that a HMM reveals longitudinal processes that should be analyzed *in addition to* traditional analyses of socio-economic factors. It is a methodological tool that reveals 'proximal causes' for a dichotomous dependent variable and, in particular, factors that collectively lead to latent states.

Using a large longitudinal sample of 4-year college students, we were able to construct cohesive states among graduates and nongraduates, which helped to follow paths toward graduation and noncompletion. This method generated important insights about the trajectories of college students, ones that cannot be easily revealed by linear modeling techniques because these do not capture the complex interactions between experiences of course-taking *across* time.

We found that graduating students rarely take STEM courses in combination with a large number of credits; instead they either withdraw from courses or they take fewer credits when attempting technical courses. Moreover, they are likely to alternate between states of 'difficult courses' and less intense states. This may be a "winning strategy" for undergraduates involved with STEM or even those enrolled in programs with a few math courses. The indication that one's STEM course-taking pattern is associated with degree completion or noncompletion has not previously appeared in the literature.

Another finding of the HMM is that noncompleters tend to start their first college semester(s) in a similar way as completers—sometimes taking many credits and taking difficult (technical) courses. They however far less frequently switch to other states in the following semesters, such as a state with a lower credit load or a state in which withdrawing courses is more common. Hence, a pattern toward noncompletion can be distinguished from the "winning strategy" by looking at the likelihood of a student to switch between states. For average American college students, academic momentum may partially emerge from the "right" course-taking pattern in the first couple of semesters. This emphasizes the importance of careful course selection early on in student's college career—a responsibility of students themselves, but also of administrators, faculty, and counselors.

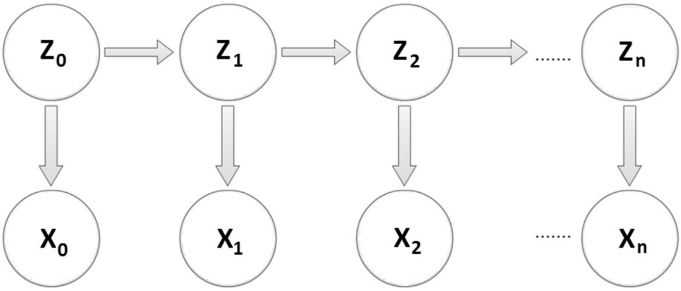
In this study, we have demonstrated the HMM's potential for prediction in educational research, as well as an extension by applying it to a sociologically relevant dependent variable in general. The predictive accuracy of an HMM with transcript data equals and in some cases outperforms logistic regression models containing socio-demographic and high school GPA variables. Moreover, these precollege variables did not predict the hidden states of the HMM. We claim that semester-by-semester observations on course-taking reflect decisions or behaviors that are not structurally associated with socio-economic background. Instead, they reveal college processes that should be analyzed *in addition to* existing socio-demographic knowledge about the graduation puzzle.

More generally, we hope that this paper has demonstrated the potential of Hidden Markov Models for analyzing sequential transcript data. Future applications could concentrate on other course and grade sequences in secondary education or community colleges, where the sequential mechanisms may be different from 4-year colleges. In addition, HMMs may be useful in examining particular subpopulations of students or majors. One application, for example, would be to take an HMM approach to STEM entry, attrition, and graduation at the college level.

Acknowledgements We thank the National Science Foundation (Grant DRL 1243785) and the Bill & Melinda Gates Foundation (Grant OPP 1012951) for their support for this study. We also thank Andrew Rosenberg (Queens College, CUNY) for his extensive technical support and his feedback on programming Hidden Markov models.

Appendix

Appendix 1: Trellis diagram of a Hidden Markov Chain



Appendix 2: Building an HMM in *Matlab*

```
O = 81;
Q = 3;
s_prior0 = normalise(rand(Q,1));
s_transmat0 = mk_stochastic(rand(Q,Q));
s_obsmat0 = mk_stochastic(rand(Q,O));
[LL_S, prior1, transmat1, obsmat1] =
dhmm_em(my_trainsuccess_data, prior0, transmat0, obsmat0,
'max_iter', 50);
```

Notes `O` = number of categories of all variables together, `Q` = number of expected states, `s_prior0` = random initial distributions, `s_transmat0` = random transition probabilities, `s_obsmat0` = random observation probabilities, `s_prior1` = expected initial distributions, `s_transmat1` = expected transition probabilities, `s_obsmat1` = expected observation probabilities, `LL_S` = log-likelihood (of the Graduation-HMM).

Appendix 3: Assessing the HMM

```
Correct = 0;
Start = 1;
End = 1045;
N = End-Start + 1;
for i = Start:End
 LL_S = dhmm_logprob(my_testfail_data (:,i), s_prior1, s_transmat1,
 s_obsmat1);
 LL_F = dhmm_logprob(my_testfail_data (:,i), f_prior1, f_transmat1,
 f_obsmat1);
 if LL_F > LL_S
  Correct = Correct + 1;
 end
end
```

Notes The ‘Start’ and ‘End’ indicate test of nongraduating students (N = 1045). The `dmm_logprob` function in the HMM toolbox was used to produce two log-likelihoods. One that matches each individual test case with the prior, transition, and observation matrices of the Graduation-HMM (`LL_S`) and one that matches each individual test case with the prior, transition, and observation matrices of the Non-Completion-HMM (`LL_F`). The `:,i` can be replaced with any selection of length of specific semester transcripts (e.g., semester 1 through 4). Finally, the algorithm classifies by comparing the log-likelihoods `LL_F` > `LL_S`

Appendix 4: Encoding and Decoding Transcripts

The input for each student-semester observation ($m_{1\ldots n}$) is based on a vector ($v_{1\ldots n}$) that has all feature values encoded using the following algorithm:

The student-semester observations (m…) include categorical and continuous values. Some categorical features include the binary features “did the student take a STEM class?” and “did the student drop any courses?” Continuous features include “number of attempted credits” and “cumulative GPA.” To simplify the modeling process, we represent all features as independent categorical features. The first step in this process is the discretization of continuous features. Each continuous value is represented as a categorical feature as described in Table 1. At this point, each vector (vi) is vector of k categorical features each of which can take one of m(k) values. The second step in the simplification process converts the vectors vi to v'i where v'i is a single categorical variable which can take one of m= \prod_{i=1}^k m^(k) values. This transformation is accomplished by a bijection, f(vi) = v'i. Since the HMM assumes that all elements in the original v vector are independent, no information is lost in via this transformation.

Subsequently, in the analysis phase, the encoded student-semester observation can be decoded through the inverse function f^-1(v'i) = v_i. Since f(v) is bijective, no information is lost in this inversion.

References

Achieve Inc. (2004). *Ready or not: Creating a high school diploma that counts. An American diploma project*. http://www.achieve.org/files/ADPreport.pdf. Accessed 24 November 2015.

Adelman, C. (1999). *Answers in the toolbox: Academic intensity, attendance patterns, and bachelor's degree attainment*. Washington, DC: U.S. Department of Education.

Adelman, C. (2004). Undergraduate grades: A complex story (Chapter 6). In C. Adelman (Ed.), *Principal indicators of student academic histories in postsecondary education* (pp. 1972–2000). Washington, DC: US Department of Education.

Adelman, C. (2006). *The toolbox revisited: paths to degree completion from high school through college*. Washington, DC: US Department of Education.

Adelman, C. (2009). *The spaces between numbers: getting international data on higher education straight*. Washington, DC: Institute for Higher Education Policy.

Aud, S., Wikinson-Flicker, S., Kristapovich P., Rathbun A., Wang X., & Zhang, J. (2013). The condition of education 2013. National Center for Education Statistics (NCES) 2013-037. Washington, DC: US Department of Education.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics, 41*(1), 164–171.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55*(4), 485–540.

Bowen, W. G., Chingos, M. M., & McPherson, M. S. (2009). *Crossing the finish line: Completing college at America's Public Universities*. Princeton, NJ: Princeton University Press.

Bozick, R. (2007). The role of students' economic resources, employment, and living arrangements. *Sociology of Education, 80*(3), 261–285.

Chen, X. (2005). *First generation students in postsecondary education: a look at their college transcripts*. National Center for Education Statistics (NCES) 2005-171. Washington, DC: US Department of Education.

Complete College America. (2011). *Time is the enemy*. Washington, DC: Complete College America. http://www.completecollege.org/docs/Time_Is_the_Enemy.pdf. Accessed 24 Nov 2015.

Duda, R. O., Hart, P. E., & Stork, D. G. (1973). *Pattern classification and scene analysis* (1st ed.). New York: John Wiley.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). New York: John Wiley.

Elzinga, C. H., Hoogendoorn, A. W., & Dijkstra, W. (2007). Linked Markov Sources modeling outcome-dependent social processes. *Sociological Methods and Research, 36*(1), 26–47.

Heck, R. H., Price, C. L., & Thomas, S. L. (2004). Tracks as emergent structures: A network analysis of student differentiation in a high school. *American Journal of Education, 110*(4), 321–353.

Hess, F., Schneider, M., Carey, K., & Kelly, A. P. (2009). *Diplomas and dropouts: Which colleges actually graduate their students (and which don't)*. Washington, DC: American Enterprise Institute.

Horn, L. & Kojaku, L.K. (2001). *High school curriculum and the persistence path through college*. National Center for Education Statistics (NCES) 2001-163. Washington, DC: US Department of Education.

Ip, E. H., Snow Jones, A., Heckert, A., Zhang, Q., & Gondolf, E. D. (2010). Latent Markov model for analyzing temporal configuration for violence profiles and trajectories in a sample of batterers. *Sociological Methods AND Research, 39*(2), 222–255.

Langeheine, R., & Van de Pol, F. (2002). Latent Markov chains. In J. Hagenaars & A. McCutcheon (Eds.), *Applied latent class analysis* (pp. 304–334). Cambridge, UK: Cambridge University Press.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley.

Murphy, K. P. (2002). *Dynamic Bayesian networks: Representation, inference and learning*. (PhD dissertation, Department of Computer Science). Berkeley, CA: University of California.

Murphy, K. P. (2005). *Hidden Markov model (HMM) Toolbox for Matlab* (Original Toolbox of 1998). http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html. Accessed 1 Aug 2016.

National Center for Education Statistics. (2011). *2004/2009 Beginning postsecondary students longitudinal study restricted use data file* [in Stata]. Washington, DC: US Department of Education, NCES 2011-244 [distributor].

Perna, L. W. (2010). Toward a more complete understanding of financial aid in promoting college enrollment. In J. Smart, *Higher education: handbook of theory and research* (Vol. 25) (pp. 129–180). New York, NY: Springer.

Perna, L. W., & Li, C. (2006). College affordability: Implications for college opportunity. *Journal of Student Financial Aid, 36*(1), 7–24.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Radford, A. W., Berkner L., Wheeless, S.C., & Shepard, B. (2011). *Persistence and attainment of 2003–2004 beginning postsecondary students: After six years*. National Center for Education Statistics (NCES) 2011-151. Washington, DC: US Department of Education.

Schneider, M., & Yin, M. L. (2012). *Completion matters: The high cost of low community college graduation rates*. Washington, DC: American Enterprise Institute for Public Policy Research.

Schuh, J. (2005). Finances and retention: Trends and potential implications. In A. Seidman (Ed.), *College student retention: Formula for student success* (pp. 277–294). Westport, CT: American Council on Education and Praeger.

Scott, S. L. (2002). Bayesian methods for hidden Markov models. *Journal of the American Statistical Association, 97*(457), 337–351.

St. John, E. P., Cabrera A. F., Nora, A., & Asker, E.H. (2000). Economic influences on persistence reconsidered. In J.M. Braxton (Ed.), *Reworking the student departure puzzle* (pp. 29–47). Nashville, TN: Vanderbilt University Press.

Stamp, M. (2015). *A revealing introduction to hidden Markov models* (Course). http://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf. Accessed 24 Nov 2015.

Tinto, V. (1993). *Leaving college: Rethinking the causes of student attrition* (2nd ed.). Chicago, IL: University of Chicago Press.

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. In S. Menard (Ed.), *Handbook of longitudinal research: design, measurement, and analysis* (pp. 373–385). Burlington, MA: Elsevier.

Wine, J., Janson, N., & Wheeless, S. (2011). *2004/09 Beginning postsecondary students longitudinal study (BPS:04/09) full-scale methodology report*. National Center for Education Statistics (NCES) 2012-246. Washington, DC: US Department of Education.

Wyner, J. S., Bridgeland, J.M., Diiulio, J. (2007). *Achievement trap: How America is failing millions of high-achieving students from lower-income families*. In V.A. Lansdowne (Ed.), Jack Kent Cooke Foundation. http://www.jkcf.org/news-knowledge/research-reports/. Accessed 24 Nov 2015.