Preliminary and Incomplete
Please do not cite

Why is the Rate of College Dropout so High and Why Is It Rising for Men?*

Alison Aughinbaugh
U.S. Bureau of Labor Statistics
Maury Gittleman
U.S. Bureau of Labor Statistics
Charles Pierret
U.S. Bureau of Labor Statistics

March 2014

* Corresponding author: Maury Gittleman, Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Room 4130, Washington, DC 20212, USA. Tel: 1 202 691-6318. Fax: 1 202 691-6310. E-mail: gittleman_m@bls.gov. This paper has been prepared for the 2014 meetings of the Population Association of America. The views expressed here are those of the authors and do not necessarily reflect the views or policies of the Bureau of Labor Statistics or any other agency of the U.S. Department of Labor.

I. Introduction

Sparked in part by the growth in the college wage premium, the proportion of high school graduates going on to post-secondary school has been on the rise in recent decades.[1] The resulting increase in the numbers of college graduates has not been proportionate, however, as the college dropout rate has gone up as well. Bound, Lovenheim and Turner (2010) estimate that the college dropout rate rose from 45.9 percent for the high school class of 1972 to 50.5 percent for the class of 1992. With a college education said to be increasingly necessary to compete in the labor market, it is important to understand why so many individuals do not achieve success in postsecondary institutions and why their numbers are growing.[2]

We address this issue by examining the college attendance and completion experience of two cohorts of the National Longitudinal Survey of Youth (NLSY), that from 1979 and that from 1997. The NLSY97 surveyed individuals who were aged 12 to 16 in 1997 every year from its inception through 2011. Information is available on every college attended, and, as a result, one can merge in a rich set of variables relating to these institutions. The same is true of the NLSY79, which has surveyed individuals who were aged 14 to 22 in 1979 over a span exceeding three decades. In this paper, we compare college completion rates between the two cohorts, and attribute differences observed to: 1) student attributes (e.g., measures of ability/achievement, college preparedness, individual demographics and family background); 2) characteristics of the postsecondary institutions (e.g., student-faculty ratios and expenditures per student) and 3) measures of any "mismatch" between the ability of the student and the quality of the institution.

[1] For instance, the proportion of high school graduates going on to college climbed from 48.4 percent for the class of 1972 to 70.7 percent for the class of 1992 (Bound, Lovenheim and Turner, 2010).

[2] The theory of the "option value" of college, that is, that one has the option but not the requirement to continue on after learning about one's abilities and preferences, predicts that some individuals will drop out (see Stange, 2012 for a recent treatment). It is an open question as to whether it can predict dropping out of a magnitude currently seen.

A mismatch is said to occur if, say, a relatively low ability student attends a relatively high quality college or if a relatively high ability student attends a relatively low quality college. Academic mismatch has proven to be empirically quite important and there is a large literature on it (see Dillon and Smith, 2013 and references therein). To our knowledge, however, no attempt has been made to explain trends in dropout rates with reference to mismatch. Light and Strayer (2000) do, however, find that mismatch plays an important role in college completion rates.

In the next section of the paper, we sketch out an empirical model of college attendance and completion. Discussion of the data is reserved for section III, which includes definitions and procedures for determining academic mismatch. Descriptive statistics are presented in section IV, while the multivariate analysis is contained in section V. Section VI offers a decomposition of changes over time in dropout rates and concluding remarks are contained in the final section. Because this paper is preliminary, the conclusion consists primarily of our plans for the next version.

II. Empirical Model of College Attendance and Completion

From the time a high school student starts thinking about colleges to which to apply to the time that student ends his/her undergraduate education, a number of decisions, many of which are sequential, have to be made. The student must first decide which colleges s/he will apply to and then from among those colleges at which s/he is accepted, must decide which ones to attend. Some will, of course, decide to attend no colleges at all. Those who attend college will be sorted into good matches, overmatches or undermatches. [3] Below, we will give precise

[3] Of course, this elides the fact that the schools have decisions about whom to accept. Dillon and Smith (2013) argue that the data are consistent with sorting into schools in terms of match group largely being the result of decisions by students not schools. That is, they conclude that most mismatches occur either because a student does not apply to a well-matched school or they apply and are admitted, but decide not to enroll.

definitions of these terms, but, for the moment, overmatches are cases where the quality of the college greatly exceeds the ability of the individual, undermatches are the opposite, and good matches are where the two quality levels are more or less closely aligned. Each year, the student, as s/he learns more information about him/herself, the school and the fit between the two must decide whether to continue in the same school, transfer to a different one, or drop out entirely. For the sake of tractability, we collapse all these decisions into two: the decision to attend college and the decision to complete college.

While the terminology in the literature and which we use as well suggests that individuals should seek "good" matches, that need not be the case, at least in theory.[4] As Dillon and Smith (2013) point out, there are a number of trade-offs that may be involved. While a low quality student at a high quality school – someone overmatched -- may face difficulties with the academic workload, s/he may benefit from being in a school with superior resources or higher quality peers. Similarly, a high quality student in a low quality school – someone who is undermatched – may be hurt by inferior resources, but may benefit from being a big fish in a small pond and consequently from receiving extra faculty attention. Further, while our school quality measures pertain to academics, there are many other dimensions along which individuals may select a school. They may wish to go to a university with strong athletic teams, where they can be with their friends, or with nice dormitories.[5] Nonetheless, we do not want to rule out that many of the mismatches are ones that individuals would have preferred to have avoided, but did not, perhaps because of financial constraints or informational barriers.

[4] In practice, Smith, Pender and Howell (2012) read the literature as suggesting that those who are undermatched are less likely to graduate from college, while Dillon and Smith (2013) assert an absence of consensus on the impact of mismatches on academic and labor market outcomes.

[5] See, for example, Jacob, McCall and Stange (2013) for evidence on demand for college consumption amenities, including spending on student activities, sports, and dormitories.

We do not present a formal theoretical model, but, instead, sketch out an informal two-period model whose empirical implementation, when complete, will be similar to that in Light and Strayer (2000), though with important conceptual differences.[6] In the first period, high school graduates, when they consider additional schooling, have four choices: 1) decide not to obtain any post-secondary education; 2) attend a school where they would be overmatched; 3) attend a school where they would be suitably matched; and 4) attend a school where they are undermatched. In the second period, those who have attended college must decide whether to complete college or to drop out.

Spelling things out further, in the first period, individuals will weigh the benefits and costs of each of the options. In addition to the non-pecuniary benefits of college attendance, each of the three college options will have a benefit of increasing earnings and providing the individual with the option of graduating college. Costs include any direct costs of attending the college, net of any financial aid received. These benefits and costs clearly differ by match group, and will be a function of both individual and school characteristics as well. In the second period, there are only two choices: dropping out or completing college. The payoff to completing college will be a function of the type of match that individuals have at the colleges they attend, in addition to other characteristics.

For the first period, we model the choice among four options as a multinomial probit. An advantage of using this form is that the error terms for each choice, which represent such things as unobserved abilities, preferences, expectations and financial costs, can be allowed to be correlated, in contrast to the case in the multinomial logit. For the second period, we also use a

[6] In the first period, the agents in Light and Strayer (2000)'s model have five choices: no college, plus attending four different types of colleges, where colleges are defined based on observed quality.

probit, but in this case a binomial one. We also allow the error term in the binomial probit to be correlated with those in the multinomial probit.[7]

III. Data

We use two nationally representative data sets to examine how well students are matched to the college that they initially attend and how the quality of the match affects their college completion, where college completion is defined as earning a bachelor's degree within eight years of completing high school. The data sets are two cohorts of the National Longitudinal Surveys of Youth (NLSY). The NLSY79 is composed of 12,686 individuals born from 1957 to 1964 and living in the United States at the start of the survey in 1979. The NLSY79 sample members ranged in age from 14 to 22 at the initial interview and on average graduated from high school and made their decisions about whether and what college to attend in 1978-1979. The NLSY97 is composed of 8,984 individuals born from 1980 to 1984 and living in the United States in 1997. The sample members of the NLSY97 ranged in age from 12 to 17 at the initial interview and on average completed high school and made their initial college decisions in 2000. Using two data sets where the sample members make decisions about college roughly 20 years apart permits us to examine how college attendance decisions and the relationship between quality of the student's match to his college and his chances of completing college may have changed over time.

In both the NLSY79 and NLSY97, information on every college attended, including name and location, is collected from the respondents. Subsequently, survey staff assign Federal Interagency Committee on Education (FICE) codes or unitids to the reported schools. These

[7] In this version of the paper, however, the error terms in the multinomial probit are independent of each other, and of the error term in the binomial probit.

codes permit us to merge the NLSY data on colleges with data from Integrated Postsecondary Education Data System (IPEDS).

We restrict the samples of analysis to those who graduate high school or earn a GED before age 21. Moreover, we consider only college attendance that begins within four years of completing high school. The definition of college completion used in this study requires the bachelor's degree to be completed within eight years of high school completion. Together these definitions mean we analyze college completion by ages 26 to 28 only.

We further restrict the samples to those who completed the Armed Services Vocational Aptitude Battery (ASVAB) as part of the NLSY, approximately 94 percent of the NLSY79 sample and 79 percent of the NLSY97 sample.[8] We use the ASVAB scores, which we discuss further below, to proxy for student ability/achievement and for how well suited the student is for the college that he attends. In addition, those respondents for whom we do not have a valid FICE/unitid code or who report a college that we are unable to match to the IPEDS data are dropped from the analysis. The economically disadvantaged non-black, non-Hispanic oversample is dropped from the NLSY79 estimation sample used.[9] The restrictions leave us with 8,728 individuals, 3,705 of whom attend two- or four-year college, in our NLSY79 sample and 6,083 individuals, 4,132 of whom attend two- or four-year college, in our NLSY97 sample.

All data are weighted by Round 1 survey weights. Appendix A provides sample definitions for all variables included in the study.

Measuring ability/achievement

To measure academic mismatch, we need measures of student quality and of college quality. For the former, we create a measure of ability/achievement using the scores on the

[8] For both the NLSY79 and NLSY97, the ASVAB subsample is a random subsample of the cohort.
[9] A disproportionate number of this sample were college students at the initial interview.

subsections of that Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB is not a measure of raw intelligence. ASVAB scores are influenced by the environment, schooling, and training that the student has had up to the point she takes the test. See Neal and Johnson (1996) for a discussion of what the ASVAB may measure.

The ASVAB was administered in both the NLSY79 and NLSY97 in order to norm new versions of the ASVAB. In the fall and summer of 1980, an ASVAB composed of ten subsections was administered to 94 percent of the NLSY79 sample. From the summer of 1997 to the spring of 1998, an ASVAB composed of twelve subsections was administered to 79 percent of the NLSY97 sample. We create measures of ability/achievement separately for the NLSY79 and NLSY97 following the same procedure.

Because the respondents were of different ages and had different amounts of schooling when they took the ASVAB, we adjust for age at administration by regressing each subsection score on quarter of birth and quarter of birth squared and then used the residuals in principal components analysis. We use the first principal component, ASVAB1, as our primary measure of ability/achievement. In the NLSY79, ASVAB1 explains 61 percent of the variance in the 10 subsections of the ASVAB. In the NLSY97, ASVAB1 explains 58 percent of the variance in the 12 subsections of the ASVAB. ASVAB1 places the highest weight on academic subjects such as general science, arithmetic reasoning, and word knowledge. The second principal component, ASVAB2, explains an additional 8 percent of the total variance in the NLSY79 and 7 percent in the NLSY97. It places the highest weight on the subsections for numerical operations and coding speed. Details are provided in Appendix B.

For both ASVAB1 and ASVAB2, we calculate the respondent's percentile within the sample distribution of high school graduates and in the sample distribution of those who attend college, in both cases weighted by the Round 1 survey weights. We use the percentile of ASVAB1 in the sample of those who go on to college to construct the measure of college mismatch. We include percentile of both ASVAB1 and ASVAB2 in the estimating equations to control for ability and achievement. The percentiles of ASVAB1 and ASVAB2 in the distribution of high school graduates are used to explain how well matched the youth is to the first college she attends. For those who start college, the percentiles of ASVAB1 and ASVAB2 in the distribution of those who attend college are used to explain the probability that the youth completes colleges with 8 years of high school completion.

Measuring college quality

In order to determine whether students are well matched to their colleges, we construct a college rank based on eight variables that are related to college quality. These include: the average salary of faculty teaching undergraduates, the student-faculty ratio, the first-to-second year retention rate, the percentage of students graduating within 150 percent of expected time (6 years for bachelor's degrees, 3 years for associate's degrees), an indicator for whether the institution had an open admissions policy, and if not, the percentage of applicants who are admitted, the percentage of admitted students who enroll, and the average SAT score (or ACT equivalent) for the freshman class. We use publically-available data from the Department of Education's Integrated Post-Secondary Education Data System's (IPEDS) Data Center for the years 2002 to 2011. While most members of the NLSY97 had already started college by 2002, we chose these years because the variables were generally only available starting in 2002. For each variable, we average values for all of these years that are available.

Following Black and Smith (2004) and Dillon and Smith (2013), our college quality variable is the first principal component across these eight variables. Relying on Black and

Smith's finding that the variables change only slowly, we use this variable for both the NLSY79 and NLSY97 data sets. To perform the match between college quality and student ability, we calculate a college's percentile rank weighted by the size of its entering cohort in 2003. Thus, "slots" for college freshman can be ranked in order. A perfect match would entail lining up the entering freshman class by ability percentile and placing them directly into colleges of equal quality percentiles. By construction, there should be approximately the same number of slots and students at each percentile ranking. See appendix C for more details.

Measuring mismatch

Following Dillon and Smith (2013), we use the difference between student ability/achievement and college quality to measure the quality of the match between the student and her college. We calculate college quality across all two- and four-year institutions included in IPEDS from years 2002 to 2011, weighted by the size of the entering cohort in 2003. We then calculate the difference between the student's location in the distribution of ASVAB1 among those who go on to college and the college's location in the distribution of college quality.

We consider students to be mismatched with their college when the difference in the percentiles exceeds 20, just as Dillon and Smith do. We refer to students as overmatched if the percentile ranking of their college exceeds their own percentile ranking by over 20 percentage points. Students are considered undermatched if their percentile ranking exceeds their college's ranking by over 20 percentage points. We refer to students who attend a college where the difference in the percentile ranking is 20 or less as having a good match, or being well-matched.[10]

To provide a better sense of how college sorting works, Table 1 displays the joint distribution between ability/achievement and college quality for college attendees in both

[10] See Smith, Pender and Howell (2012) for an alternative approach relying on Barron's categorization.

cohorts. In a 4 x 4 grid with 16 cells, with no sorting at all, each cell would have 6.25 percent of the overall weight and each column and row percentage would equal 25 percent. If everyone sorted into an exact match –that is, percentile of student quality equals percentile of college quality -- all the weight would be on the diagonal elements, so that each of these four cells would have 25 percent of the overall weight and account for 100 percent of the column and row weights. An examination of Table 1 reveals that there clearly is a substantial amount of sorting, but not so much that there is not plenty of scope for mismatch. For both cohorts, about half of those in the 4th quartile (top) in ability/achievement are in the 4th (top) quartile of school quality, and about one-third of those in the 1st (bottom) quartile in ability/achievement are in the corresponding quartile of school quality. But for the other two ability quartiles, the diagonal elements are not the most populated ones. The second ability/achievement quartile is fairly evenly distributed across schooling quality quartiles, particularly for NLSY79, while the highest concentration of 3rd quartile ability/achievement individuals is in the 4th quartile of school quality.

IV. Descriptive Statistics

We begin the descriptive analysis with a comparison of college attendance between the two cohorts. As shown in Table 2, there was a dramatic rise in the share of high school graduates going on to college between the two the cohorts, from 46.83 percent to 70.45 percent, which is similar to the increase estimated by Bound, Lovenheim and Turner (2010). While the gain was large for both sexes, consistent with the story of women catching up to and then surpassing men in college attendance (Goldin, Katz and Kuziemko, 2006), the increase was greater for females (48.42 percent to 75.93 percent) than for males (45.22 percent to 65.02 percent). Thus, women have opened up a gap of about 11 percentage points for the 1997 cohort.

The proportion of those attending college now starting at 2-year schools as opposed to four-year ones rose rapidly, from 41.15 to 53.09 percent. The increase was more rapid among men than women, about 15 percentage points (38.53 to 53.64 percent), versus from 11 percentage points (43.57 to 52.61 percent) for women. With the majority of the cohort now beginning college life at a 2-year school, trends at such institutions will obviously have a major impact on the overall picture. Given that some community college students are seeking sub-baccalaureate vocational training, a question arises as to whether it is appropriate to include them in an analysis of 4-year college completion rates. Bound, Lovenheim and Turner (2010), in a related study, argue that it is for two reasons: First, they estimate that some 70 percent of community college entrants in a cohort they examine intend to complete a bachelor's degree. Second, as attendance at a community college provides an option to continue to the completion of a four-year degree, these students are important in the determination of cohort college completion rates.

While these arguments are compelling, especially because the share of those starting college at a 2-year school who intend to obtain a bachelor's degree may have shifted over time, we think it is important to examine the sensitivity of our results to the implicit assumption made by Bound, Lovenheim and Turner, that all entrants into a 2-year college are potentially interested in the option of completing a 4-year degree. Thus, we will perform the analysis both including and excluding those who start college at 2-year institutions.

Because of our interest in assessing the role played by academic mismatch, among other factors, in college completion trends, we also show on Table 2 the distribution of college students by their mismatch status. A few points are worthy of mention. As noted by others, the degree of mismatch is substantial.[11] First, for the 1979 cohort, while about half of the cohort had a good match, some 22 percent were overmatched, and about 30 percent were undermatched. For the 1997 cohort, about half continued to have a good match, while 17 percent were overmatched and 36 percent were undermatched. Second, for the 1979 cohort, men were much more likely to be undermatched than women (37 percent to 23 percent), both because they were less likely to have a good match and because they were less likely to be overmatched. For the 1997 cohort, the distribution across the match groups looks much more similar by sex, with women becoming less likely to overmatch and more likely to undermatch. Third, when comparing the distributions by 4-year colleges and 2-year colleges, it is readily apparent that students are much more likely to be overmatched in 4-year colleges than in 2-year colleges and more likely to be undermatched in 2-year colleges than in 4-year colleges. This situation results from the fact that the 2-year colleges are at the bottom of the college rankings. Thus, the increase in the proportion of students at 2-year colleges is a key reason why the overall distribution has shifted away from overmatch and towards undermatch.

Table 3 offers a look at how college completion rates have changed between the cohorts. As the share of the cohorts going on to college grew, the proportion of those who started college who actually completed it shows little change when both sexes are combined, staying at 39 percent. The stability in this number, however, masks opposing trends by sex. The share of men completing college declined by about 4 percentage points from 40.90 percent to 36.83 percent, while that for women rose by almost exactly the same amount from 37.71 percent to 41.70 percent. Thus, the completion rate for women now exceeds that of men, despite the fact that a greater share of female high school graduates go on to college.

[11] As we use a similar approach to Dillon and Smith (2013) and they also use the NLSY97, it is no surprise that we have similar amounts of mismatch for this dataset. Smith, Pender and Howell (2012) focus on undermatch and estimate that some 41 percent of the class of 2004 was undermatched, down from 49 percent for the class of 1992.

An alternative perspective is gained by examining completion rates by mismatch group and by mismatch group interacted with sex. For the full sample in NLSY79, completion rates for those overmatched and those with a good match are fairly close (44 percent versus 40 percent), while that for those undermatched is much lower at 34 percent. Relative to the corresponding match group in that cohort, those who were overmatched in NLSY97 showed a substantial 13 percent increase in completion rates. The completion rate for good matches changed little, while that for those undermatched declined by 6 percent. The net impact of these changes is that, for the full sample in NLSY97, completion rates are ordered by implied college quality. That is, given an individual's ability, the highest quality colleges will be overmatches, the next best in the good match category, and the remaining will be undermatches. It is in that order that completion rates are ranked in the full sample for the two sexes combined, and for men and women separately.

This strengthening in the completion rate between cohorts noted for those overmatched was experienced by both sexes, but it was more than offset for men by the decline in completion rates both among those who have good matches and among those who are undermatched. For women, however, there was an increase in the completion rate among those who had a good match and only a slight decline for the undermatched.

When completion rates are broken down by whether the initial school was a 2-year or a 4-year one, a number of differences by school type become readily apparent. First, for those who start at 2-year schools, for both cohorts, those who undermatch have the highest completion rates. For those who began at 4-year colleges in NLSY79, good matches have the highest completion rates, followed by undermatches and then by overmatches. But in the NLSY97 cohort, a sharp increase in the completion rate among those who are overmatched and a decline among those who are undermatched has reversed the positions of these two groups. The category good matches maintains the highest completion rate, however.

Second, it is probably not surprising that 4-year completion rates in all categories are much higher for those who start at 4-year schools than for those who start at 2-year schools. It still seems worthy of note, however, that the gap between the two widened from 37 percentage points to 48 percentage points between the two cohorts. Finally, it is clear that the decline in completion rates among men is ascribable to the situation at 2-year colleges, as completion rates actually rose at 4-year schools (by 7.67 percentage points), though by a slower rate than that for women (12.51 percentage points). For those starting at 2-year schools, the rate of completion for men actually declined by 4.58 percentage points versus an increase of 3.15 percentage points for women.

How do individual and college characteristics vary by match quality? Have these relationships changed over time? We take a first pass at these questions by displaying, in Table 4, the mean values of basic demographics, family background, characteristics at time graduated high school, ability/achievement as well as college characteristics, by cohort and match group. Of particular interest are differences across the match groups and differences between the cohorts. The results by sex highlight what we noted earlier, that women are more likely than men to overmatch and less likely than men to undermatch, but the gap in the latter narrowed over time. By race/ethnicity, blacks are more likely to overmatch than other groups, perhaps reflecting affirmative action, but the relative likelihood of their overmatching has declined over time. The relationship between income and overmatching appears to have changed between cohorts as well. For the 1979 cohort, the first or bottom quartile was the only one whose representation in overmatch exceeded its share of the total population, and only slightly. For the

1997 cohort, perhaps reflecting a growing importance of financial resources, the top income quartile is clearly overrepresented among overmatches. Turning to ability/achievement, those characteristics that are used to defined student quality (ASVAB1) will, for arithmetic reasons, have means that increase as one goes from overmatch to good match to undermatch. That is, because the higher an individual's ASVAB1 is the lower the probability that there will be a school that that individual can be overmatched with. Variables that are highly correlated will also follow the same pattern, which the SAT score does. Interestingly, the high school GPA does in 1979, but goes in the opposite direction in 1997. We will have more to say on this below. Percentile in college quality distribution, once again, has a mechanical relationship to match group, as it is more likely that higher quality schools will be involved in overmatches and lower quality ones in undermatches. For the 1979 cohort, average expenditures per student decline as one heads toward undermatches, but for the 1997 cohort, the three categories are tightly bunched. For average faculty salary, there is a fairly tight bunching for 1979, but there is a clearer ordering from overmatches to good matches to undermatches in 1997. For student-faculty ratio, there is not much difference across match groups in 1979, but for 1997, this ratio is clearly higher for the undermatches.

V. Multivariate Analysis

A. College Attendance

Tables 5 and 6 summarizes result from four multinomial probit models, the first two from NLSY79 and the second two from NLSY97. There are two models on each table. In the first of the pair, the choices are: 1) to not attend any college; 2) to attend a college where the individual will be overmatched; 3) to attend a college where the individual will have a good match; and 4) to attend a college where the individual will be undermatched. In the second of the pair, the choices are about 4-year colleges only so that the choices are: 1) to not attend any 4-year college; 2) to attend a 4-year college where the individual will be overmatched; 3) to attend a 4-year college where the individual will have a good match; and 4) to attend a 4-year college where the individual will be undermatched. Marginal effects are reported for the four outcomes. Normalizations for the multinomial probit require that one set of a coefficients, those for a base outcome, be set to zero. In estimation, our base outcome was a good match. Though it is possible to calculate marginal effects for all four outcomes and we present all four, in the discussion that follows we pay less attention to those for a good match, given that they were derived from the coefficients for the other three.

For the first model for NLSY79, summarized on Table 5, ASVAB1 has a strong positive relationship with going to college. As noted previously, its relationship with the three match outcomes is more mechanical, however. The higher is ASVAB1, the fewer schools there exist for an overmatch, making it not surprising that ASVAB1's coefficient for this outcome is negative and statistically significant. By similar reasoning, it is not surprising that ASVAB1's coefficient for undermatch is positive and statistically significant. For other ability/achievement variables, ASVAB2 is a significant predictor of college attendance and of overmatch, and high school GPA is a predictor just of college attendance. Taking SAT appears to be more relevant than actual SAT in predicting college attendance, but it also has a positive relationship with both overmatching and undermatching.

With other controls, women are still less likely to be overmatched and more likely to be undermatched, as noted above. All else equal, both Hispanics and Blacks are more likely than whites to attend college, and they are also more likely than whites to overmatch. The higher are father's and mother's education levels, the more likely it is the child will go on to college. These two variables are also positive predictors of overmatch.

The second model, which combines in one outcome 2-year starters with those who do not attend college, provides a closer look at predictors at match type at 4-year schools. Once again, ASVAB1 has a significant negative relationship with overmatching and a significant positive one with undermatching. ASVAB2 is a positive predictor of overmatch, as is the indicator for having taken the SAT. Women continue to be less likely to overmatch, all else equal, but there is no significant difference in undermatching. Whites are less likely to overmatch. As before, both father's and mother's education levels are positive predictors of overmatch, but now father's education has a negative relationship to undermatch.

Turning to the later cohort, on Table 6 one can see a very similar pattern in terms of significant marginal effects as in Table 5. For the first model, one key difference is for sex. For NLSY79, sex was not a significant predictor of college attendance, but did have a negative relationship with overmatch and a positive one with undermatch. For NLSY97, women, all else equal, are now more likely to attend college and are no longer less likely to overmatch. They do, however, continue to be more apt to undermatch. In terms of race, Blacks and Hispanics continue to be more likely to attend college than whites, but they are no longer more likely to overmatch. For the second model, the results tend to be similar to those for NLSY79, but with some differences as just noted for sex and race.

B. College Completion

Table 7 presents results from six probit models of college completion estimated on our NLSY79 sample. Panel A uses the full sample, while Panel B includes only those who started at baccalaureate institutions. As we find important differences by sex, we show results from the same models separately by sex.

The mismatch variables have large effects in NLSY79, especially among all college attendees where those overmatched have an estimated 14 percentage point higher probability of completing than those with a good match, while those with an undermatch have a 15 percentage point lower probability. Among four-year college attendees, the match effects have the same ordering, but the marginal effect of being overmatched is 9 percentage and that of being undermatched is -10 percentage points. Thus, for both samples college completion rates are ordered in terms of implicit college quality. While, as shown on Table 3, women were about 3 percentage points behind men in college completion rates for the full NLSY79 sample and actually a little bit ahead among 4-year starters, when individual and school characteristics are controlled for, the probability of a woman graduating is about 9 percentage points lower than a man.

Measures of ability/achievement such as ASVAB1, ASVAB2, SAT score and GPA are all significant positive predictors of whether an individual completes college among the full sample, although the SAT score is not significant when the sample is restricted to 4-year starters. Among family background variables, father's grade completed has a significant positive effect. Those from the highest income quartile appear to be at an advantage among the full sample, but that is not the case in the restricted sample. Those who get financial aid are less likely to graduate in the full sample, perhaps because it is an indicator of resource constraints, but there is no significant relationship in the latter sample.

Finally, while we are implicitly measuring the impact of school resources through our mismatch variables, we seek to capture direct effects by examining the influence of average expenditures per student, mean faculty salaries and student-faculty ratios. Only the last mentioned is statistically significant and of the right sign, with higher ratios being associated with a lower probability of graduation for both samples.

Given the large differences by sex just noted, it is useful to examine the estimates separately by men and women. For the full sample, the results by sex are broadly similar, though father's grade completed and financial aid receipt are significant among men and not among women. On the other hand, lower student-faculty ratios seem to help women complete college, but not men. Among 4-year starters, for men, overmatching does not provide a significant benefit, as undermatching continues to provide an advantage, while for women it is the opposite. Other differences by sex are that father's grade completed is only relevant for men and income quartile and student-faculty ratio only for women.

In Table 8, we report results for NLSY97 for the same six probits as we did for NLSY79 in Table 7. Given the attention we just paid to differences by sex, it is first of interest to note that after controlling for other characteristics, women in the full sample of college attendees are predicted to have very similar completion rates to men, in contrast to case in NLSY97, when they trailed by 9 percentage points. The differences by match group have widened. For the full sample, those who are overmatched have an estimated 17 percentage point advantage over those who have good matches, while those who have good matches have an edge of 18 points over those who are undermatched. Other results look more similar between the cohorts. Both ASVAB1 and ASVAB2 are significant predictors of college completion, as is high school GPA. Family background variables seem somewhat more important for the later cohort, as both parents' education variables are significant, in contrast to just the father's. Moreover, the positive relationship between income and graduation appears to be a bit stronger as well. Financial aid is a significant negative predictor of graduating in both cases.

The results for the college characteristics for the full sample are somewhat discrepant across the cohorts. Faculty salaries have a positive association with graduation in NLSY97, while they were insignificant in NLSY79. Expenditures are of the wrong sign (negative) and significant in the later cohort, but not significant in the earlier one. Student faculty-ratios are not significant in NLSY97, while they had a significant negative association in NLSY79.

Examining the results separately by sex one tends not to see major differences. Perhaps of greatest interest are the results for college characteristics. For men, only the expenditure variable is significant, but it has the wrong sign. For women, all three college resource variables are significant, but the marginal effect for expenditures again has the wrong sign.

After restricting the sample to those who started in 4-year schools, we re-examine the determinants of college completion. Once again, the marginal effect for sex is not significant. By race/ethnicity, Hispanics are less likely to finish college than are whites, a relationship that had not been significant among all college attendees. Only father's highest grade completed and not mother's highest grade completed is statistically significant. Among college characteristics, the expenditure variables is no longer significant, but the other relationships remain intact.

Examining the results separately by sex, we once again focus on the college characteristics. For this sample, the results are more similar, as for both sexes, only the salary variable is statistically significant.

VI. Decomposition Analysis

Results from decomposition analysis are presented in Table 9. The observed completion rates are shown in the top two rows, followed by the difference in the third row. The fourth row shows the difference between the observed NLSY79 rate and the counterfactual completion rate calculated using all observables from the NLSY97 data. Differences in the distribution of the observables do not explain the difference in completion rates between the two cohorts. With the exception of the model estimated for men who begin college at a 4-year institution, the residuals are at least as large as the change due to the observables, implying that the relationships between the explanatory variables and college completion rates changed over this period as well.

We use a method developed by Fairlie (2005) that extends Blinder-Oaxaca decompositions to nonlinear models in order to examine the effects of changes in student characteristics versus changes in college characteristics. The remainder of the table presents the predicted change in completion rates based on these decompositions. The simulations show that student characteristics changed in ways that raise college completion rates while college characteristics changed in ways that decrease the probability of completing college.

Changing the distribution of match quality from that observed in the NLSY79 to that of the NLSY97 does not change completion rates. Changes in other student characteristics, however, do lead to increases in the college completion rate. Because a greater percentage of high school graduates go on to college in the NLSY97 compared to the NLSY79 and the additional entrants, on average, have lower ability/achievement, the simulations show that substituting the distribution of achievement/ability for the NLSY97 college attendees into the model predicts a two to three percentage point decrease in the NLSY79 college completion rate. Because most of the additional entrants into college start at 2-year institutions, the effect of changing in the distribution of ability/achievement for those begin at baccalaureate granting school is smaller by about a one percentage point.

Changing from the NLSY79 to the NLSY97 distributions for the level of college preparedness (measured by high school GPA, whether a student took the SAT/ACT, and scores on the SAT/ACT) and for parents' education increase the probability that students complete college. The simulations of changes in college preparation show large increases in completion rates. In all cases, the predicted increase from improvements in college preparation exceed the predicted decrease from lower ability/achievement. Among those who begin college at 4-year institutions, the effect from increases in college preparation are four to six times larger than the decrease from lower ability/achievement.

As shown by Bound, Lovenheim, and Turner (2010) average college characteristics changed over time in ways that lowered the probability of completing college. In general, all college characteristics that we control for work in the same direction, with the largest decreases coming from a change in the distribution of student-faculty ratios. The NLSY79 completion rate falls by about 3 percentage points when the distribution of student-faculty ratio is replaced by that in the NLSY97.

VII. Concluding Remarks

In order to look at the questions of why college dropout rates are high and why the completion rates among men are falling in a convincing way, we have additional work to do. To begin with the data need to be augmented in several ways, including incorporating college costs, family wealth, and local geographic information into the estimation.

Second, we need to consider whether our results are sensitive to the way that we measure the quality of match between the student and her college and quality of the college. While we use the same college rankings in the two data sets, we treat the student's ability/achievement differently in that we measure student ability relatively among those who attend college in each of the data sets. Consequently, a high school student in the 60th percentile in the ASVAB1 distribution of high school graduates would likely be in the bottom quartile of the ASVAB1 distribution for those who attend college in the NLSY79 where approximately 47 percent of high school graduates go on to college, but probably near the middle of the distribution in the NLSY97 where approximately 70 percent of high school graduates attend college. The fact that the current match definition permits different values across the two cohorts for a pair composed of the same institution and a student with same level of ability/achievement among high school graduates seems curious. It may be more appropriate to control for student ability and college quality separately and to control additionally for the interaction between the two.

Third, the current estimation is naïve. The current model does not permit the outcome of not attending college versus the outcomes of attending a college where the student is overmatched, a good match, or under matched to be correlated. Moreover, in the current estimation the students' attendance decision and her completion decision are treated as independent. As we continue to work on this project, we plan to move away from this naïve model and instead estimate a model which permits the student's decisions to be correlated.

While this is a work in progress, it is clear that part of the explanation for why the college dropout rate is high is that many more students are attending 2-year colleges in the recent cohort. While the completion rate is approximately equal in the NLSY79 and NLSY79 among those who begin college at either a 2-year or a 4-year college, the completion rates for those who start at a 4-year college have risen substantially from 55 percent to 65 percent.

References

Black, Dan and Jeffrey Smith (2004): “How Robust is the Evidence on the Effects of College Quality? Evidence from Matching,” *Journal of Econometrics*, 121(1), pp. 99-124.

Bound, John, Michael F. Lovenheim, and Sarah Turner. “Why Have College Completion Rates Declined? An Analysis of Changing Student Preparation and Collegiate Resources.” *American Economic Journal: Applied Economics* 2 (July 2010): 129–157.

Dillon, Eleanor Wiske and Jeffrey Andrew Smith. “The Determinants of Mismatch Between Students and Colleges.” NBER Working Paper No. 19286, August 2013.

Fairlie, Robert W. "An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models," *Journal of Economic and Social Measurement*, 30(4): 305-316.

Goldin, Claudia, Lawrence F. Katz, and Ilyana Kuziemko. “The Homecoming of American College Women: The Reversal of the College Gender Gap.” *Journal of Economic Perspectives*, Volume 20, Number 4, Fall 2006, pp. 133–156.

Jacob, Brian, Brian McCall and Kevin Stange. 2013. “College as Country Club: Do Colleges Cater to Students’ Preferences for Consumption?” unpublished, University of Michigan, January 17.

Light, Audrey and Wayne Strayer. “Determinants of College Completion: School Quality or Student Ability?” *Journal of Human Resources*, Vol. 35, No. 2 (Spring, 2000), pp. 299-332.

Neal, Derek A. and William R. Johnson "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy*, 1996, *104*(5), pp. 869-95.

Smith, Jonathan, Matea Pender and Jessica Howell (2012): “The Full Extent of Student-College Academic Undermatch,” Unpublished manuscript, the College Board, October.

Stange, Kevin M. “An Empirical Investigation of the Option Value of College Enrollment.” *American Economic Journal: Applied Economics* 2012 4 (1): 49–84.

Table 1: Joint Distribution of Ability and College Quality

	College Quality										
	NLSY79					NLSY97					
Ability	1	2	3	4	Total	1	2	3	4	Total	
1	513	383	351	154	1,401	438	410	344	106	1,298	frequency
	36.64	27.94	24.38	11.03	100.00	33.22	32.35	24.93	9.50	100.00	weighted row percentage
	38.63	31.85	25.49	9.06	25.00	34.81	33.04	26.54	8.45	25.02	weighted column percentage
	9.16	6.99	6.10	2.76	25.00	8.31	8.09	6.24	2.38	25.02	weighted cell percentage
2	229	230	215	192	866	281	280	272	201	1,034	
	25.54	26.94	24.20	23.31	100.00	26.51	27.30	25.72	20.46	100.00	
	26.97	30.76	25.34	19.18	25.04	27.75	27.86	27.36	18.18	24.99	
	6.39	6.75	6.06	5.84	25.04	6.63	6.82	6.43	5.11	24.99	
3	157	147	191	260	755	190	214	220	310	934	
	19.70	18.00	25.30	37.00	100.00	20.26	22.72	23.06	33.96	100.00	
	20.76	20.50	26.44	30.37	24.99	21.21	23.18	24.52	30.16	24.99	
	4.92	4.50	6.32	9.24	24.99	5.06	5.68	5.76	8.49	24.99	
4	97	117	151	318	672	133	140	177	416	866	
	12.96	14.84	21.77	50.44	100.00	15.51	15.59	20.28	48.62	100.00	
	13.65	16.89	22.74	41.39	24.98	16.24	15.91	21.58	43.21	25.00	
	3.24	3.71	5.44	12.60	24.98	3.88	3.90	5.07	12.16	25.00	
Total	996	877	908	924	3705	1042	1044	1013	1033	4132	
	23.71	21.93	23.91	30.44	100.00	23.88	24.49	23.50	28.13	100.00	
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	
	23.71	21.93	23.91	30.44	100.00	23.88	24.49	23.50	28.13	100.00	

Table 2: College attendance by type of college and match quality

	NLSY79 (n=8728)			NLSY97 (n=6083)		
	Total	Men	Women	Total	Men	Women
Percent of high school graduates attending college	46.83	45.22	48.42	70.45	65.02	75.93
Distribution by type of college						
4-year colleges	58.85	61.47	56.43	46.91	46.36	47.39
2-year colleges	41.15	38.53	43.57	53.09	53.64	52.61
Distribution by quality of match with college						
Overmatch	21.60	16.49	26.33	17.17	14.61	19.38
Good match	48.37	46.32	50.27	47.12	45.65	48.40
Undermatch	30.04	37.20	23.41	35.71	39.74	32.22
4-year colleges						
Overmatch	34.41	25.12	43.79	31.62	26.54	35.94
Good match	49.54	52.03	47.02	51.95	52.87	51.17
Undermatch	16.05	22.85	9.19	16.43	20.59	12.90
2-year colleges						
Overmatch	3.26	2.72	3.71	4.40	4.31	4.47
Good match	46.69	37.19	54.47	42.85	39.40	45.90
Undermatch	50.04	60.09	41.82	52.75	56.29	49.63

Note: Data are weighted by Round 1 weights. The sample size for the NLSY79 is 8728 and for the NLSY97 is 6083. The number who attend college in each cohort is 3705 in the NLSY79 and 4132 in the NLSY97.

Table 3: College completion rates by match quality

	NLSY79				NLSY97				Difference			
	Total	Overmatch	Good match	Undermatch	Total	Overmatch	Good match	Undermatch	Total	Overmatch	Good match	Undermatch
Full Sample	39.24	44.20	40.42	33.77	39.44	56.81	41.82	27.94	0.20	12.61	1.40	-5.83
Men	40.90	39.24	45.87	35.44	36.83	53.20	40.55	26.55	-4.07	13.96	-5.32	-8.89
Women	37.71	47.09	35.78	31.31	41.70	59.17	42.86	29.43	3.99	12.08	7.08	-1.88
4-year	54.55	46.46	60.35	53.97	64.87	64.66	69.21	51.57	10.32	18.20	8.86	-2.40
Men	54.04	41.59	60.50	53.01	61.71	62.52	66.34	48.77	7.67	20.93	5.84	-4.24
Women	55.05	49.27	60.18	56.37	67.56	66.01	71.73	55.37	12.51	16.74	11.55	-1.00
2-year	17.36	10.21	10.19	24.51	16.96	6.88	12.48	21.44	-0.40	-3.33	2.29	-3.07
Men	19.92	4.51	13.19	24.78	15.34	3.64	10.64	19.52	-4.58	-0.87	-2.55	-5.26
Women	15.25	13.63	8.51	24.18	18.40	9.65	13.89	23.36	3.15	-3.98	5.38	-0.82

Note: Data are weighted by Round 1 weights. Sample size for NLSY79 is 3705 and for NLSY97 is 4132.

Table 4: Average characteristics by match quality

	NLSY79				NLSY97			
	Total	Overmatch	Good match	Undermatch	Total	Overmatch	Good match	Undermatch
Basic Demographics								
Gender								
Male	0.48	0.37	0.46	0.60	0.46	0.40	0.47	0.52
Female	0.52	0.63	0.54	0.40	0.54	0.60	0.53	0.48
Race/Ethnicity								
Black, non-Hispanic	0.13	0.27	0.13	0.03	0.12	0.18	0.13	0.07
Hispanic	0.05	0.07	0.06	0.03	0.10	0.11	0.11	0.08
non-Black, non-Hispanic	0.82	0.66	0.81	0.94	0.76	0.69	0.74	0.83
Family Background								
Number of Siblings	2.93	3.10	2.98	2.74	3.36	3.59	3.36	3.24
Mother's high grade	12.57	12.60	12.52	12.63	13.70	13.79	13.67	13.68
Father's high grade	13.12	13.13	13.25	13.11	13.83	13.84	13.90	13.74
Characteristics at time graduated high school								
Region								
NE	0.21	0.29	0.21	0.15	0.18	0.27	0.18	0.15
MW	0.29	0.22	0.30	0.33	0.28	0.24	0.27	0.30
S	0.32	0.38	0.31	0.29	0.32	0.30	0.32	0.33
W	0.18	0.11	0.18	0.23	0.22	0.19	0.23	0.22
Urban	0.82	0.85	0.79	0.83	0.70	0.73	0.72	0.67
Family income by quartile								
First quartile	0.14	0.15	0.13	0.13	0.13	0.16	0.13	0.12
Second quartile	0.15	0.13	0.16	0.15	0.15	0.13	0.15	0.16
Third quartile	0.17	0.15	0.16	0.19	0.19	0.18	0.20	0.18
Fourth quartile	0.23	0.23	0.24	0.21	0.20	0.25	0.22	0.16

Missing	0.31	0.34	0.31	0.32	0.33	0.28	0.30	0.38
Proportion of college educated					0.25	0.27	0.25	0.24
High school student-teacher ratio								
<14	0.09	0.09	0.09	0.07	0.35	0.41	0.34	0.32
14-18	0.27	0.25	0.29	0.26	0.33	0.29	0.34	0.33
18-22	0.44	0.47	0.43	0.46	0.19	0.20	0.19	0.21
>22	0.20	0.19	0.21	0.21	0.13	0.10	0.14	0.14
Ability/Achievement								
Percentile in ability1 distribution for college attendees	50.48	41.34	48.18	69.93	50.49	30.16	45.04	67.43
Percentile in ability2 distribution for college attendees	50.48	66.42	51.42	41.12	50.49	56.10	51.40	46.58
High school GPA	2.82	2.64	2.83	2.92	3.07	3.11	3.07	3.05
Whether took SAT/ACT	0.42	0.44	0.40	0.44	0.59	0.68	0.56	0.57
SAT score	905.89	791.90	922.15	960.87	1,032.67	989.19	1,039.82	1,047.44
College characteristics								
Percentile in college quality distribution	45.91	66.42	48.18	26.74	40.15	67.36	44.48	21.36
Two-year	0.41	0.06	0.40	0.69	0.53	0.14	0.48	0.78
Four-year	0.59	0.94	0.60	0.31	0.47	0.86	0.52	0.22
Ave. expenditures per student (in 1982-84$)	8,282.64	8,768.39	8,430.04	7,694.71	8,706.04	8,680.53	8,858.01	8,535.17
Ave. faculty salary (in 1982-84$)	80,946.28	82,062.96	81,614.90	79,061.03	74,438.22	82,279.02	75,872.98	68,698.95
Student-faculty ratio	18.57	17.32	18.40	19.73	31.74	23.11	26.81	42.56
Sample Size	3,705	972	1,798	935	4,132	786	2,009	1,337

Note: Data are weighted by Round 1 weights. Sample size for NLSY79 is 3705 and for NLSY97 is 4132.

Table 5: Multinomial Probit of College Attendance and Match, NLSY79
Marginal Effects

5A: Match type includes 2-year colleges

	No College		Overmatch		Good Match		Undermatch	
Female	0.004		-0.019	*	-0.009		0.024	*
Hispanic	-0.192	**	0.060	**	0.131	**	0.000	
Black	-0.157	**	0.072	**	0.096	**	-0.011	
ASVAB1	-0.004	**	-0.002	**	0.000		0.006	**
ASVAB2	-0.002	**	0.002	**	0.001	**	0.000	
High school GPA	-0.057	**	0.010		0.057	**	-0.01	
SAT	0.000		0.000		0.000	*	0.000	
Took SAT	-0.232	**	0.121	**	0.023		0.088	**
Father’s highest grade completed	-0.020	**	0.009	**	0.014	**	-0.003	
Mother’s highest grade completed	-0.018	**	0.011	**	0.008	**	0.000	
Family income quartile 2	-0.010		-0.018		0.034		-0.006	
Family income quartile 3	-0.004		-0.012		0.014		0.002	
Family income quartile 4	-0.059	**	0.017		0.059	**	-0.017	
Urban	-0.049	**	0.021	*	-0.004		0.032	**
HS student-teacher ratio 1	0.032		0.008		0.014		-0.053	**
HS student-teacher ratio 2	0.029		-0.001		0.005		-0.033	*
HS student-teacher ratio 3	0.012		-0.002		-0.004		-0.006	
HS student-teacher ratio 4	-0.002		-0.018		-0.001		0.021	
Home resources	0.035		-0.011		-0.017		-0.007	
Home resources missing	0.031		0.029		-0.071		0.011	

Notes: All data are weighted by Round 1 weights. The sample sizes are 8728 for the full sample, 4355 for men, and 4373 for women. * indicates p<0.05, and ** p<0.01. Additional regressors include year of birth, whether high school transcript was collected, whether family income was missing, whether urbanicity was missing, and whether HS student-teacher ratio was missing.

5B. Match type excludes 2-year colleges

	No College or 2-year College		Overmatch		Good Match		Undermatch	
Female	0.048	**	-0.021	*	-0.021		-0.006	
Hispanic	-0.111	**	0.055	**	0.048	**	0.008	
Black	-0.166	**	0.063	**	0.078	**	0.025	*
ASVAB1	-0.004	**	-0.002	**	0.003	**	0.003	**
ASVAB2	-0.003	**	0.002	**	0.001	**	0	
High school GPA	-0.062	**	0.009		0.041	**	0.012	*
SAT	0		0		0		0	
Took SAT	-0.187	**	0.105	**	0.059	*	0.024	
Father's highest grade completed	-0.015	**	0.008	**	0.009	**	-0.002	*
Mother's highest grade completed	-0.018	**	0.011	**	0.006	**	0.001	
Family income quartile 2	-0.02		-0.021		0.028		0.013	
Family income quartile 3	0.003		-0.014		-0.003		0.013	
Family income quartile 4	-0.055	**	0.012		0.039	**	0.004	
Urban	-0.005		0.019	*	-0.009		-0.005	
HS student-teacher ratio 1	0.01		0.014		-0.001		-0.023	*
HS student-teacher ratio 2	0.001		0.005		0.012		-0.018	*
HS student-teacher ratio 3	0.016		-0.001		0.004		-0.019	*
HS student-teacher ratio 4	0.065	**	-0.016		-0.021		-0.028	**
Home resources	0.007		-0.02		-0.036		0.049	
Home resources missing	-0.005		0.027		-0.071		0.049	

Notes: All data are weighted by Round 1 weights. The sample sizes are 8728 for the full sample, 4355 for men, and 4373 for women. * indicates p<0.05, and ** p<0.01. Additional regressors include year of birth, whether high school transcript was collected, whether family income was missing, whether urbanicity was missing, and whether HS student-teacher ratio was missing.

Table 6: Multinomial Probit of College Attendance and Match, NLSY97
Marginal Effects

6A: Match type includes 2-year colleges

	0=No college		Overmatch		Good Match		Undermatch	
Female	-0.058	**	0.009		0.019		0.030	*
Black	-0.050	**	0.019		0.036		-0.005	
Hispanic	-0.071	**	0.013		0.041	*	0.016	
ASVAB1	-0.003	**	-0.003	**	-0.002	**	0.008	**
ASVAB2	-0.002	**	0.001	**	0.001	**	0.000	
High school GPA	-0.104	**	0.056	**	0.108	**	-0.060	**
SAT	0.000		0.000	**	0.000	*	0.000	**
Took SAT	-0.130	**	0.010		-0.053		0.172	**
Father's highest grade completed	-0.013	**	0.004	*	0.014	**	-0.006	*
Mother's highest grade completed	-0.017	**	0.008	**	0.009	**	0.000	
Family income quartile 2	-0.004		-0.007		0.023		-0.012	
Family income quartile 3	-0.060	**	0.005		0.069	**	-0.014	
Family income quartile 4	-0.095	**	0.041	**	0.086	**	-0.031	
Urban1	-0.035	**	0.006		0.030	*	-0.001	
Urban2	0.007		-0.009		-0.021		0.023	
HS student-teacher ratio 1	-0.085	**	0.032	*	0.054	*	-0.001	
HS student-teacher ratio 2	-0.080	**	0.012		0.062	**	0.007	
HS student-teacher ratio 3	-0.078	**	0.028		0.031		0.020	
HS student-teacher ratio 4	-0.112	**	0.010		0.066	*	0.036	
Home resources	-0.049		0.000		0.043		0.006	
Home resources missing	0.005		-0.046		0.024		0.017	

Notes: All data are weighted by Round 1 weights. The sample sizes are 6083 for the full sample, 3028 for men, and 3055 for women. * indicates p<0.05, and ** p<0.01. Additional regressors include year of birth, whether high school transcript was collected, whether family income was missing, whether urbanicity was missing, and whether HS student-teacher ratio was missing.

6B: Match type excludes 2-year colleges

	0=No college or 2-year College		Overmatch		Good Match	Undermatch	
Female	-0.015		0.012		0.001	0.002	
Black	-0.092	**	0.027	*	0.045	0.019	
Hispanic	0.005		0.014		-0.021	0.002	
ASVAB1	-0.003	**	-0.002	**	0.003	0.003	**
ASVAB2	-0.001	**	0.001	**	0.001	0.000	
High school GPA	-0.148	**	0.051	**	0.084	0.014	
SAT	0.000		0.000	**	0.000	0.000	
Took SAT	-0.069		0.002		0.064	0.003	
Father's highest grade completed	-0.013	**	0.004	*	0.010	-0.001	
Mother's highest grade completed	-0.013	**	0.008	**	0.005	0.000	
Family income quartile 2	-0.025		0.006		0.030	-0.011	
Family income quartile 3	-0.074	**	0.019		0.054	0.001	
Family income quartile 4	-0.108	**	0.059	**	0.051	-0.002	
Urban1	-0.007		0.004		0.011	-0.009	
Urban2	0.058		-0.018		-0.011	-0.028	
HS student-teacher ratio 1	-0.069	**	0.048	**	0.022	-0.001	
HS student-teacher ratio 2	-0.048	*	0.023		0.019	0.007	
HS student-teacher ratio 3	-0.026		0.035	*	-0.010	0.000	
HS student-teacher ratio 4	0.015		0.014		-0.027	-0.002	
Home resources	-0.035		-0.002		-0.021	0.058	
Home resources missing	0.014		-0.043		-0.035	0.064	

Notes: All data are weighted by Round 1 weights. The sample sizes are 6083 for the full sample, 3028 for men, and 3055 for women. * indicates p<0.05, and ** p<0.01. Additional regressors include year of birth, whether high school transcript was collected, whether family income was missing, whether urbanicity was missing, and whether HS student-teacher ratio was missing.

Table 7: College Completion, NLSY79
Marginal Effects

7A: Sample of all college attendees

	All College Attendees		Men		Women	
Female	-0.094	**				
Hispanic	-0.010		-0.013		-0.003	
Black	0.026		0.018		0.040	
Overmatch	0.143	**	0.080	*	0.167	**
Undermatch	-0.148	**	-0.150	**	-0.160	**
ASVAB1-college	0.005	**	0.005	**	0.006	**
ASVAB2-college	0.003	**	0.004	**	0.002	**
High school GPA	0.128	**	0.132	**	0.113	**
SAT	0.000	*	0.000		0.000	
Took SAT	-0.036		-0.076		0.004	
Father's highest grade completed	0.008	**	0.011	*	0.006	
Mother's highest grade completed	0.007		0.005		0.009	
Number of siblings	-0.006		-0.004		-0.008	
Family income quartile 2	0.028		0.006		0.050	
Family income quartile 3	0.054		0.038		0.067	
Family income quartile 4	0.067	*	0.056		0.074	
Northeast	0.046		0.050		0.059	
Midwest	0.018		0.069		-0.009	
South	0.019		0.081		-0.012	
West	-0.060		-0.040		-0.054	
Urban	-0.009		0.003		-0.035	
HS student-teacher ratio 1	0.038		0.078		0.021	
HS student-teacher ratio 2	0.031		0.080	*	-0.004	
HS student-teacher ratio 3	-0.016		-0.002		-0.025	
HS student-teacher ratio 4	-0.028		-0.026		-0.034	
Financial aid	-0.148	*	-0.224	*	-0.040	
Ln(Average expenditure per student)	0.004		-0.001		0.011	
Ln(Average faculty salary)	-0.016		-0.081		0.011	
Student-faculty ratio	-0.004	*	-0.002		-0.006	**
College resources missing	-0.225		-0.897		0.026	

Notes: All data are weighted by Round 1 weights. The sample sizes are 4132 for all college attendees, 1879 for men and 2253 for women. * indicates $p<0.05$, and ** $p<0.01$. Additional regressors include year of birth, whether high school transcript was collected, whether family income was missing, whether region was missing, whether urbanicity was missing, whether HS student-teacher ratio was missing, and whether financial aid was missing. Dollar amounts are converted to 1983-1984 dollars using the CPI-U.

7B: Sample of those who begin college at a 4-year college

	4-year starters only		Men		Women	
Female	-0.085	**				
Hispanic	-0.048		-0.026		-0.053	
Black	0.033		0.055		0.030	
Overmatch	0.088	**	0.048		0.108	*
Undermatch	-0.100	**	-0.123	**	-0.105	
ASVAB1-college	0.006	**	0.005	**	0.006	**
ASVAB2-college	0.003	**	0.004	**	0.003	**
High school GPA	0.140	**	0.139	**	0.134	**
SAT	0.000		0.000		0.000	
Took SAT	-0.044		-0.103		0.007	
Father's highest grade completed	0.010	*	0.017	**	0.004	
Mother's highest grade completed	0.006		-0.002		0.011	
Number of siblings	-0.005		0.000		-0.014	
Family income quartile 2	-0.007		-0.018		0.023	
Family income quartile 3	0.035		-0.026		0.095	
Family income quartile 4	0.061		0.000		0.126	*
Northeast	0.077		0.123		0.045	
Midwest	-0.019		0.065		-0.084	
South	0.009		0.086		-0.041	
West	-0.062		0.012		-0.110	
Urban	0.017		0.029		-0.013	
HS student-teacher ratio 1	0.046		0.095		0.030	
HS student-teacher ratio 2	0.003		0.055		-0.027	
HS student-teacher ratio 3	-0.034		0.001		-0.056	
HS student-teacher ratio 4	-0.035		-0.027		-0.052	
Financial aid	-0.042		-0.046		0.050	
Ln(Average expenditure per student)	-0.056		-0.062		-0.058	
Ln(Average faculty salary)	-0.041		-0.102		0.037	
Student-faculty ratio	-0.007	*	-0.003		-0.011	**
College resources missing	-0.577		-1.129		0.159	

Notes: All data are weighted by Round 1 weights. The sample sizes are 1852 for all 4-year starters, 842 for men, and 1010 for women. * indicates p<0.05, and ** p<0.01. Additional regressors include year of birth, whether high school transcript was collected, whether family income was missing, whether region was missing, whether urbanicity was missing, whether HS student-teacher ratio was missing, and whether financial aid was missing. Dollar amounts are converted to 1983-1984 dollars using the CPI-U.

Table 8: College Completion, NLSY97
Marginal Effects

8A: Sample of all college attendees

	All College Attendees		Men		Women	
Female	0.010					
Black	-0.015		0.008		-0.032	
Hispanic	-0.038		-0.034		-0.041	
Overmatch	0.172	**	0.192	**	0.161	**
Undermatch	-0.184	**	-0.175	**	-0.190	**
ASVAB1-college	0.004	**	0.004	**	0.004	**
ASVAB2-college	0.001	**	0.001	**	0.002	**
High school GPA	0.154	**	0.178	**	0.129	**
SAT	0.000		0.000	*	0.000	
Took SAT	-0.002		-0.028		0.013	
Father's highest grade completed	0.010	**	0.008		0.013	**
Mother's highest grade completed	0.009	**	0.009	*	0.009	
Number of siblings	-0.005		-0.003		-0.005	
Family income quartile 2	0.048		0.030		0.064	
Family income quartile 3	0.118	**	0.111	**	0.113	**
Family income quartile 4	0.115	**	0.121	**	0.103	**
Northeast	-0.142	**	-0.194	**	-0.057	
Midwest	-0.151	**	-0.204	**	-0.062	
South	-0.163	**	-0.187	**	-0.099	
West	-0.146	**	-0.225	**	-0.037	
Urban	0.010		0.008		0.011	
HS student-teacher ratio 1	0.106	**	0.091	*	0.117	**
HS student-teacher ratio 2	0.113	**	0.106	**	0.116	**
HS student-teacher ratio 3	0.076	**	0.040		0.102	*
HS student-teacher ratio 4	0.096	**	0.106	*	0.089	
Financial aid	-0.097	**	-0.112	**	-0.083	**
Ln(Average expenditure per student)	-0.069	**	-0.097	**	-0.061	*
Ln(Average faculty salary)	0.016	*	0.018		0.017	*
Student-faculty ratio	0.000		0.000		-0.001	*
College resources missing	-0.617	**	-0.871	**	-0.534	

Notes: All data are weighted by Round 1 weights. The sample sizes are 3705 for all college attendees, 1731 for men and 1974 for women. * indicates p<0.05, and ** p<0.01. Additional regressors include year of birth, whether high school transcript was collected, whether family income was missing, whether region was missing, whether urbanicity was missing, whether HS student-teacher ratio was missing, and whether financial aid was missing. Dollar amounts are converted to 1983-1984 dollars using the CPI-U.

8B: Sample of those who begin college at a 4-year college

	4-year Starters Only		Men		Women	
Female	0.015					
Black	-0.035		-0.022		-0.041	
Hispanic	-0.093	*	-0.073		-0.115	*
Overmatch	0.127	**	0.150	**	0.109	**
Undermatch	-0.188	**	-0.175	**	-0.176	**
ASVAB1-college	0.003	**	0.002	*	0.003	**
ASVAB2-college	0.001	**	0.001		0.002	**
High school GPA	0.138	**	0.211	**	0.063	
SAT	0.000		0.000		0.000	
Took SAT	-0.088		-0.104		-0.066	
Father's highest grade completed	0.011	*	0.006		0.013	*
Mother's highest grade completed	0.006		0.011		0.002	
Number of siblings	0.001		0.005		0.000	
Family income quartile 2	0.029		0.012		0.052	
Family income quartile 3	0.127	**	0.143	*	0.113	*
Family income quartile 4	0.089	*	0.126	*	0.061	
Northeast	-0.164	*	-0.129		-0.307	*
Midwest	-0.199	**	-0.156		-0.336	*
South	-0.193	**	-0.136		-0.348	**
West	-0.139		-0.103		-0.273	*
Urban	-0.010		-0.021		-0.001	
HS student-teacher ratio 1	0.057		0.073		0.048	
HS student-teacher ratio 2	0.089		0.084		0.089	
HS student-teacher ratio 3	0.043		0.021		0.054	
HS student-teacher ratio 4	0.066		0.072		0.050	
Financial aid	-0.064		-0.107	*	-0.017	
Ln(Average expenditure per student)	-0.073		-0.100		-0.047	
Ln(Average faculty salary)	0.036	**	0.059	*	0.036	**
Student-faculty ratio	0.000	*	0.000		-0.002	
College resources missing	-0.635		-0.893		-0.396	

Notes: All data are weighted by Round 1 weights. The sample sizes are 2047 for all 4-year starters, 1005 for men, and 1042 for women. * indicates p<0.05, and ** p<0.01. Additional regressors include year of birth, whether high school transcript was collected, whether family income was missing, whether region was missing, whether urbanicity was missing, whether HS student-teacher ratio was missing, and whether financial aid was missing. Dollar amounts are converted to 1983-1984 dollars using the CPI-U

Table 9: Decomposition of change in completion rates based on NLSY79 estimates, by type of institution and gender

	Any college			4-year college		
	All	Men	Women	All	Men	Women
NLSY79	39.24	40.90	37.71	54.55	54.04	55.05
NLSY97	39.44	36.83	41.70	64.87	61.07	67.56
Total change	0.19	-4.06	3.99	10.32	7.67	12.51
Change due to observables	2.11	2.32	0.59	4.49	10.65	-0.16
Change due to student characteristics	6.07	3.37	6.81	6.78	8.74	5.96
Match quality	-0.70	-0.07	-1.45	-0.20	0.28	-1.00
Ability/Achievement	-3.31	-3.41	-2.22	-1.06	-1.09	-0.75
College Preparation	3.72	3.61	3.57	4.92	5.28	4.18
Parents' Education	0.98	0.81	1.13	1.28	1.08	1.62
High school characteristics	1.12	2.27	0.58	1.61	2.78	1.19
Demographics	-0.07	0.38	0.07	0.28	1.22	-0.22
Other student characteristics	4.33	-0.22	5.14	-0.04	-0.82	0.95
Change due to college characteristics	-3.86	-1.17	-6.26	-2.32	1.92	-6.12
Student-faculty ratio	-2.68	-1.68	-3.70	-2.89	-1.20	-4.38
Ln(expenditures per student)	-0.47	-0.34	-0.70	-0.09	-0.02	-0.18
Ln(faculty salary)	-0.83	0.15	-1.78	0.22	0.53	-0.11
Missing--any college characteristic	0.11	0.69	-0.10	0.44	2.65	-1.45
Residual	-1.92	-6.38	3.40	5.83	-2.98	12.67

Notes: All data weighted by Round 1 weights. Ability/achievement includes ASVAB1 and ASVAB2. College preparation includes high school GPA, whether took SAT/ACT, and SAT/ACT score. Parents’ education includes both father’s highest grade and mother’s highest grade. Demographics includes race/ethnicity, sex, and year of birth. Other student characteristics include region, urbanicity, number of siblings, family income, and whether the student received financial aid.

Appendix A: Variable definitions

Variable	NLSY79	NLSY97
Dependent Variables		
Match	Quality of match between student and the college she attends. Equal to 0 if she does not attend college, equal to 1 if overmatch, 2 if good match, 3 if undermatch. Overmatch (undermatch) indicates student's percentile in ASVAB1 distribution of college attenders is less (greater) than the college's percentile in the distribution of college quality by over 20 points.	Quality of match between student and the college she attends. Equal to 0 if she does not attend college, equal to 1 if overmatch, 2 if good match, 3 if undermatch. Overmatch (undermatch) indicates student's percentile in ASVAB1 distribution of college attenders is less (greater) than the college's percentile in the distribution of college quality by over 20 points.
Complete college	Earns a bachelor's degree within 8 years of completing high school	Earns a bachelor's degree within 8 years of completing high school
Independent variables in match equation		
Sex	Dummy variable equal to 1 if respondent is female	Dummy variable equal to 1 if respondent is female
Race/ethnicity	Series of dummy variables indicating respondent is Hispanic, Black, or non-black and non-Hispanic. Non-black, non-Hispanic is the omitted category.	Series of dummy variables indicating respondent is Hispanic, Black, of mixed race, or non-black and non-Hispanic. Non-black, non-Hispanic is the omitted category.
Year of Birth	Series of dummy variable indicating year of birth. 1957 is omitted year.	Series of dummy variable indicating year of birth. 1980 is omitted year.
ASVAB1	Percentile rank in first principal component of 10 ASVAB subsection scores among high school completers	Percentile rank in first principal component of 12 ASVAB subsection scores among high school completers
ASVAB2	Percentile rank in second principal component of 10 ASVAB subsection scores among high school completers	Percentile rank in second principal component of 12 ASVAB subsection scores among high school completers
High school GPA	High school GPA calculated from high school transcript	High school GPA calculated by survey staff from high school transcript
SAT	SAT score from high school transcript; ACT scores are converted to a SAT equivalent	SAT score from high school transcript; ACT scores are converted to a SAT equivalent
Took SAT	Indicator that student took SAT or ACT	Indicator that student took SAT or ACT
Transcript missing	Indicator that high school transcript data is not available as part of NLSY79	Indicator that high school transcript data is not available as part of NLSY97
Father's highest grade completed	Continuous variable ranging from 1 to 20	Continuous variable ranging from 1 to 20, use biological father first and

		residential father if data on bio father is not available
Mother's highest grade completed	Continuous variable ranging from 1 to 20	Continuous variable ranging from 1 to 20, use biological mother first and residential father if data on bio mother is not available
Father's highest grade completed missing	Indicator that father's highest grade is missing. In this case, father's highest grade is set to zero.	Indicator that father's highest grade is missing. In this case, father's highest grade is set to zero.
Mother's highest grade completed missing	Indicator that mother's highest grade is missing. In this case, mother's highest grade is set to zero.	Indicator that mother's highest grade is missing. In this case, mother's highest grade is set to zero.
Quartile of household income in year respondent graduated high school	Series of dummy variables indicating quartile of household income in the year the respondent graduated high school from the distribution of the NLSY79. Fifth category indicates household income was missing. Lowest quartile of income is the omitted category.	Series of dummy variables indicating quartile of household income in the year the respondent graduated high school from the distribution of the NLSY79. Fifth category indicates household income was missing. Lowest quartile of income is the omitted category.
Urban	Dummy variable indicating that respondent lives in urban area in year s/he graduates from high school. Set to 0 if missing.	Dummy variable indicating that respondent lives in urban area in year s/he graduates from high school. Set to 0 if missing.
Urban missing	Dummy variable indicating that information on whether respondent lives in an urban area is missing.	Dummy variable indicating that information on whether respondent lives in an urban area is missing.
High school, student-teacher ratio	Series of dummy variables indicating the student-teacher ratio at the student's high school. Categories are for missing, less than 14, 14 to less than 18, 18 to less than 22, 22 and higher. From school survey.	Series of dummy variables indicating the student-teacher ratio at the student's high school. Categories are for missing, less than 14, 14 to less than 18, 18 to less than 22, 22 and higher. Merge to school survey data from QED.
Home resources	Indicator that respondent's home received newspaper, magazine, or had a library card at age 14. Zero if missing.	Indicator that student had home computer or quiet place to study at age 14. Zero if missing.
Home resources missing	Indicator that some of the variables that make up the home resources variable are missing. In this case, home resources is set to zero.	Indicator that some of the variables that make up the home resources variable are missing. In this case, home resources is set to zero.
Independent variables in college completion equation		
ASVAB1-college	Percentile rank in first principal component of 10 ASVAB subsection scores among those who attend college	Percentile rank in first principal component of 12 ASVAB subsection scores among those who attend college
ASVAB2-college	Percentile rank in second principal	Percentile rank in second principal

	component of 10 ASVAB subsection scores among those who attend college	component of 12 ASVAB subsection scores among those who attend college
Match	Series of dummy variables indicating undermatch and overmatch between student and his/her college. Being well-matched to college is the omitted category.	Series of dummy variables indicating undermatch and overmatch between student and his/her college. Being well-matched to college is the omitted category.
Sex	Dummy variable equal to 1 if respondent is female	Dummy variable equal to 1 if respondent is female
Race/ethnicity	Series of dummy variables indicating respondent is Hispanic, Black, or non-black and non-Hispanic. Non-black, non-Hispanic is the omitted category.	Series of dummy variables indicating respondent is Hispanic, Black, of mixed race, or non-black and non-Hispanic. Non-black, non-Hispanic is the omitted category.
Year of Birth	Series of dummy variable indicating year of birth. 1957 is omitted year.	Series of dummy variable indicating year of birth. 1980 is omitted year.
High school GPA	High school GPA calculated from high school transcript	High school GPA calculated by survey staff from high school transcript
SAT	SAT score from high school transcript; ACT scores are converted to a SAT equivalent	SAT score from high school transcript; ACT scores are converted to a SAT equivalent
Took SAT	Indicator that student took SAT or ACT	Indicator that student took SAT or ACT
Transcript missing	Indicator that high school transcript data is not available as part of NLSY79	Indicator that high school transcript data is not available as part of NLSY97
Father's highest grade completed	Continuous variable ranging from 1 to 20	Continuous variable ranging from 1 to 20, use biological father first and residential father if data on bio father is not available
Mother's highest grade completed	Continuous variable ranging from 1 to 20	Continuous variable ranging from 1 to 20, use biological mother first and residential father if data on bio mother is not available
Father's highest grade completed missing	Indicator that father's highest grade is missing. In this case, father's highest grade is set to zero.	Indicator that father's highest grade is missing. In this case, father's highest grade is set to zero.
Mother's highest grade completed missing	Indicator that mother's highest grade is missing. In this case, mother's highest grade is set to zero.	Indicator that mother's highest grade is missing. In this case, mother's highest grade is set to zero.
Quartile of household income in year respondent graduated high school	Series of dummy variables indicating quartile of household income in the year the respondent graduated high school from the distribution of the NLSY79. Fifth category indicates household income was missing. Lowest quartile of income is the omitted category.	Series of dummy variables indicating quartile of household income in the year the respondent graduated high school from the distribution of the NLSY79. Fifth category indicates household income was missing. Lowest quartile of income is the omitted category.

Region	Dummy variable indicating region in which respondent lived in year s/he graduated from high school. Categories for missing, Northeast, Mid-west, South, and West.	
Urban	Dummy variable indicating that respondent lives in urban area in year s/he graduates from high school. Set to 0 if missing.	Dummy variable indicating that respondent lives in urban area in year s/he graduates from high school. Set to 0 if missing.
Urban missing	Dummy variable indicating that information on whether respondent lives in an urban area is missing.	Dummy variable indicating that information on whether respondent lives in an urban area is missing.
Financial aid	Indicator that student received scholarship, grant, or loan to attend college. Set to zero if missing. Self-reported	Indicator that student received scholarship, grant, work-study or loan to attend college. Set to zero if missing. Self-reported.
Financial aid missing	Dummy variable indicating that financial aid information is missing.	
Ln(Average faculty salary)	Logarithm of average faculty salary at student's college from 1980 IPEDS data. In 1982$. Set to zero if missing.	Logarithm of average faculty salary at student's college from 2001 IPEDS data. In 1982$. Set to zero if missing.
Ln(Average expenditure per student)	Logarithm of average expenditures per student at student's college from 1980 IPEDS data. In 1982$. Set to zero if missing.	Logarithm of average faculty salary at student's college from 1999-2000 IPEDS data. In 1982$. Set to zero if missing.
Student-faculty ratio	Student-faculty ratio at student's college form 1980 IPEDS data. Set to zero if missing.	Student-faculty ratio at student's college form 2000 IPEDS data. Set to zero if missing.
College resources missing	Dummy variable equal to one if any of the college characteristics from IPEDS are missing.	Dummy variable equal to one if any of the college characteristics from IPEDS are missing.

Appendix B: Principal components of the ASVAB sub-tests in the NLSY79 and NLSY97

Scores on each subsection of the ASVAB are adjusted for the age of the respondents when they took the test by regressing the score on the quarter of birth and quarter of birth squared and using the residuals for the principal component analysis. The first two principal components combined explain 69.0 percent of the total variance of the 10 test subsection scores in the NLSY79 and 65.0 percent of the total variance of the 12 test subsection scores in the NLSY97.

A. NLSY79

	First component	Second component	Unexplained variance
Eigenvalue	6.080	0.818	
Total variance explained	0.608	0.082	
Eigenvectors:			
General science	0.866	-0.133	0.233
Arithmetic reasoning	0.851	0.060	0.272
Word knowledge	0.870	0.070	0.239
Paragraph comprehension	0.800	0.187	0.324
Numerical operations	0.669	0.418	0.378
Coding Speed	0.595	0.425	0.466
Auto information	0.720	-0.418	0.308
Mathematics knowledge	0.791	0.161	0.349
Mechanical comprehension	0.781	-0.316	0.290
Electronics information	0.810	-0.319	0.243

B. NLSY97

	First component	Second component	Unexplained variance
Eigenvalue	6.918	0.882	
Total variance explained	0.577	0.074	
Eigenvectors:			
General science	0.865	-0.141	0.232
Arithmetic reasoning	0.853	0.145	0.252
Word knowledge	0.856	-0.041	0.266
Paragraph comprehension	0.839	0.123	0.281
Numerical operations	0.628	0.423	0.434
Coding Speed	0.582	0.366	0.528
Auto information	0.606	-0.367	0.498
Shop information	0.643	-0.439	0.394
Mathematics knowledge	0.848	0.282	0.201
Mechanical comprehension	0.808	-0.192	0.310
Electronics information	0.801	-0.241	0.301
Assembling objects	0.698	0.099	0.503

Appendix C –Construction of College Quality Indicator using Principal Components

Our college quality variable is constructed as the first principal component of an analysis using eight variables: the average salary of faculty teaching undergraduates, the student-faculty ratio, the first-to-second year retention rate, the percentage of students graduating within 150 percent of expected time (6 years for bachelor's degrees, 3 years for associate's degrees), an indicator for whether the institution had an open admissions policy, and if not, the percentage of applicants who are admitted, the percentage of admitted students who enroll, and the average SAT score (or ACT equivalent) for the freshman class. If the school uses open enrollment, acceptance rate is assumed to be 100%, SAT Scores are set at 400, and the percentage of admitted students enrolling is set at 60%. As the table below shows, the first component puts little weight on faculty salaries or student-faculty ratio, and only moderate weight on the percentage of accepted students accepting. All other components are heavily weighted with the expected signs.

	First component	Second component	Unexplained variance
Eigenvalue	3.520	1.331	
Total variance explained	0.440	0.166	
Eigenvectors:			
Faculty Salaries	0.146	0.776	0.377
Student-Faculty Ratio	0.053	0.787	0.378
Retention Rate - 1st to 2nd year	0.811	-0.066	0.338
Graduation Rate	0.749	-0.229	0.386
Open Admissions	-0.793	0.074	0.366
Acceptance Rate	-0.799	0.039	0.360
Percent of Accepted Students Enrolling	-0.487	-0.204	0.721
SAT Score	0.878	0.069	0.224