NATIONAL
CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

High Schools and Students' Initial Colleges and Majors

Rajeev Darolia
Cory Koedel

WORKING PAPER 165 • April 2018

High Schools and Students' Initial Colleges and Majors

Rajeev Darolia
University of Missouri

Cory Koedel
University of Missouri/CALDER

Contents

Acknowledgements

We thank the Missouri Department of Higher Education for providing access to data and gratefully acknowledge research support from the National Center for Analysis of Longitudinal Data in Education Research (CALDER) funded through grant #R305C120008 to American Institutes for Research from the Institute of Education Sciences, U.S. Department of Education. We thank session participants at Purdue University and the APPAM, CALDER, and AEFP conferences for valuable feedback, in particular Mark Long and Carrie Conaway. The usual disclaimers apply.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

High Schools and Students' Initial Colleges and Majors
Rajeev Darolia, Cory Koedel
CALDER Working Paper No. 165
April 2018

Abstract

We use statewide administrative data from Missouri to examine the explanatory power of high schools over student sorting to colleges and majors at 4-year public universities. We develop a "preparation and persistence index" (PPI) for each university-by-major cell in the Missouri system that captures dimensions of selectivity and rigor and allows for a detailed investigation of sorting. Our analysis shows that students' high schools predict the quality of the initial university, as measured by PPI, conditional on their own academic preparation, and that students from lower-SES high schools systematically enroll at lower-PPI universities. However, high schools offer little explanatory power over major placements within universities.

1 Introduction

College and major placements play an important role in shaping students' academic and post-college outcomes. These placements also collectively influence the human capital of the workforce, which is important in light of concerns that students in the United States are no longer keeping pace with their global competitors in developing the key skills that promote long-term economic prosperity (Committee on Prospering in the Global Economy of the 21st Century, 2007). For these reasons, and because the socioeconomic backgrounds of students are unequally distributed across universities and majors, recent research has focused increasingly on the factors that explain how and why students enroll in different colleges and pursue different majors (Arcidiacono, Aucejo, and Hotz, 2016; Bowen, Chingos and MacPherson, 2009; Hoxby and Turner, 2014; Hurwitz et al., 2017; Porter and Umbach, 2006; Stinebrickner and Stinebrickner, 2014; Wiswall and Zafar, 2015).

We contribute to the literature on college and major sorting by examining the role of high schools in explaining students' initial university and major placements conditional on postsecondary enrollment. To facilitate our investigation of sorting we develop a new, empirical measure to quantify student sorting across university-by-major cells in the Missouri state university system. Our measure is based on the pre-college academic qualifications of students who complete a degree in each cell, where the qualifications are weighted based on how well they predict student success in college.[1] We refer to our new measure as the "preparation and persistence index" (PPI).

Variation in PPI across university-by-major cells arises from differences in admissions decisions, students' initial choices, persistence within cells, and cross-cell transfers. Thus, it captures dimensions of selectivity and rigor at the university-by-major level. The PPI is more flexible and differentiated than metrics that are commonly used to track student placements in college. Several conceptual benefits derive from the flexible, empirical foundation of PPI. For example, PPI facilitates rankings of majors that overlap across universities when the universities differ by the overall level of selectivity. It also allows us to move

[1] Our preferred measure of success is graduation from college within 8 years, but our findings are qualitatively similar if we use other college outcomes (see below).

away from traditional, subjective divisions of college majors such as between STEM and non-STEM majors, and relatedly, allows for a better accounting of heterogeneity within groups of traditionally-defined STEM and non-STEM fields (also see Webber, 2016).

We document the across- and within-university variance shares of cell-level PPI in the Missouri system. Universities explain a substantial fraction of the total variance of PPI – about 62 percent – but the within-university variance is substantial as well (38 percent). We also explore related variability in the academic alignment between students and their entering university-by-major cells. This analysis complements previous research focusing on academic "undermatching" of students to university placements (Arcidiacono and Lovenheim, 2016; Dillon and Smith, 2017; Hoxby and Turner, 2014; Smith, Pender, and Howell, 2013), which we extend to consider placements of students to majors within universities. This investigation is motivated by evidence that college and major selectivity, and the interaction, explain labor market returns to education (Eide, Hillmer, and Showalter, 2015; Thomas and Zhang, 2005; Webber, 2016).

Turning to our analysis of high schools, a number of studies examine how high schools influence academic performance in college. Previous research has focused on outcomes such as college grades, persistence, and graduation (e.g., Betts and Morrell, 1999; Black, Lincove, Cullinane, and Veron, 2015; Fletcher, 2012; Fletcher and Tienda, 2010; Fletcher and Mayer, 2013; Long, Iatarola, and Conger, 2009). Our contribution is to examine the predictive power of high schools over students' initial university-by-major placements. We report on the overarching predictive power of high schools, inclusive of the influence of the communities in which they are situated, as well as the predictive power of selected observed high-school and local-area characteristics. Our dataset is well-suited to investigate the mapping from high schools to university-by-majors cells because we observe large numbers of students who enter and exit the Missouri university system via various college and major pathways from hundreds of high schools in the state.

We show that high schools are strong predictors of entering-cell PPI conditional on students' own academic preparation. This result is driven primarily by the explanatory power of high schools over university placements. Consistent with previous research (e.g., Dillon and Smith, 2017; Hoxby and Turner, 2014; Smith, Pender, and Howell, 2013), our preferred specifications indicate that students from lower-SES high schools systematically enroll at lower-PPI universities relative to their similarly-prepared peers from higher-SES high schools. We also extend this line of inquiry to examine sorting within universities. Despite the presence of substantial variation in the PPI of entering-major cells within universities, high schools explain a negligible fraction of the variance in students' within-university placements.

2 Context and Data

We use administrative microdata provided by the Missouri Department of Higher Education (DHE) for the empirical analysis. We focus our attention on six cohorts of full-time, state-resident, non-transfer students who entered the public 4-year university system in Missouri from a public high school between 1996 and 2001 as college freshman. Because inclusion in our dataset requires initial enrollment at a 4-year public university, our analysis is not informative about college-attendance outcomes. Instead, we focus on students' university and major placements conditional on enrollment. In total, our analytic sample includes 58,377 students. Basic descriptive statistics are provided in Appendix Table A.1.[2]

We identify collegiate major pathways based on the Classification of Instructional Programs (CIP) taxonomy developed by the US Department of Education.[3] We define majors as specific to each university. This means that we treat students who enter the same major (i.e., same CIP code) at different universities as entering via separate pathways. We also note that in Missouri, like in other states, university enrollment is not entirely separable from major enrollment because universities have different major offerings. In total,

[2] Our dataset is similar to the dataset used by Arcidiacono and Koedel (2014). Notable differences between the datasets are that we include students from all racial and ethnic groups in our data, whereas they restrict their analysis to African American and white students, and we restrict our attention to students who matriculate into the system from public high schools.

[3] We aggregate majors at the 4-digit CIP code level. For sparsely populated university-by-major cells (those with less than 10 who start or less than 5 that finish in the cell), we aggregate them with other majors within the 2-digit CIP code level, but this type of aggregation affects a small number of students (approximately four percent of completers obtain a degree with a CIP code that must be aggregated).

over the course of our data panel we identify 476 unique university-by-major cells in the Missouri 4-year public university system.

The initial major that we use to define the entering cell is best interpreted as an "intended" major because there are no requirements or formal system rules that govern the initial selection (e.g., a student can declare herself to be a business major upon entry, prior to being officially accepted into the business program). Though not formally binding, the initial major is important because it shapes students' initial plans of study, peers, and advisors.[4] We match enrollment data to completion records to identify a final university and major for each graduate. Each student is tracked for eight years to determine graduation outcomes; all individuals who do not obtain a degree within eight years from a university in the Missouri system are coded as non-completers.[5]

We observe students' high schools of attendance and for many high schools we observe large numbers of students entering the 4-year university system.[6] Thus, our data are well-suited to examine the transition from high schools to university-by-major cells, given that we typically have large unit-level samples at both levels. The DHE data additionally include detailed information on the pre-college academic preparation of individual students – most notably, students' class percentile ranks and ACT scores. We use these data to (a) construct the empirically-derived PPIs for each university-by-major cell as described in the next section, and (b) investigate the role of high schools in determining student sorting conditional on students' own pre-entry academic preparation. Again, we use "high school" to denote the high school itself and the surrounding area.

The degree of student sorting to public universities in Missouri will be less than the degree of sorting to universities more broadly given the scope of heterogeneity among postsecondary institutions

---

[4] Furthermore, as documented below, the initial major is highly predictive of the final major. In cases where students list multiple majors, we identify the primary major based on the first listed major.
[5] In robustness analyses, we use of measures of graduation in four and six years and find similar results.
[6] We drop records from approximately 3 percent of in-state students who do not have an assigned high school of attendance in the DHE data or who come from high schools that send a small number (<10) of students to an in-state, public university during the period. We observe students who attended 455 different public high schools.

nationally and internationally, and in the public and private sectors. Nonetheless, there is substantial heterogeneity across the 13 public 4-year universities in the state system, mapped in Figure 1.[7] The University of Missouri-Columbia is the flagship university and only university with the highest research activity distinction. The other highly selective universities are Truman State University and the STEM-focused Missouri University of Science and Technology.[8] There are also two historically black universities in the system, Harris-Stowe State University and Lincoln University (the latter is a land grant university).[9]

We provide additional information about Missouri universities in Table 1. The universities are ordered by the average of an individual academic preparation index for entering students in the first column (we describe the preparation index in the next section). There are several notable features of the system. Beginning with how enrollment is distributed across universities, the third column shows that over forty percent of students in the analytic sample enter into just two universities: the University of Missouri-Columbia and Missouri State University. No other university has more than a 10-percent enrollment share. Variation in the index also tends to be the least among the universities with the highest average pre-entry preparation indices.

The fourth column of Table 1 shows the eight-year graduation rate for each campus (determined by tracking students in our sample for up to eight years after entry to see if a bachelor's degree was obtained). Graduation rates map fairly closely to the pre-entry preparation index in column 1. The most notable differences occur at the urban campuses, University of Missouri-Kansas City and University of Missouri-St. Louis, which have lower graduation rates than would be predicted by students' pre-entry preparation alone. The low graduation rates at the urban campuses are consistent with similar results

[7] We use the word "system" to describe all 13 Missouri universities. In terms of governance, there are several subsystems of universities (e.g., the 4-campus "University of Missouri" system) but we do not distinguish between these subsystems in our work.
[8] Based on the 2015 Carnegie Classifications of Higher Education. See http://carnegieclassifications.iu.edu. We use the term "highly selective" to characterize institutions with an undergraduate profile considered "more selective" in the Carnegie lexicon (the highest level of selectivity).
[9] The HBCUs may generate a different type of sorting. The findings from our analysis of high schools are not generally sensitive to whether we include students who attend the HBCUs in our analytic sample. However, HBCU enrollment does systematically lower university-level placements for students who attend high schools with large minority enrollment shares, all else equal, which is not surprising given that the HBCUs are low-ranked by PPI per Table 1 and disproportionately attended by minority students (as shown by Arcidiacono and Koedel, 2014).

reported using Missouri data in Arcidiacono and Koedel (2014), and more broadly for urban campuses in Bowen, Chingos and McPherson (2009), who show that graduation rates are negatively related to the commuter share.

Finally, the last two columns of Table 1 display the average and standard deviation of the academic preparation index among graduates. As expected, the average index is higher among graduates than non-graduates, which can be seen by comparing the inclusive index values in column 1 with the graduate-only values in column 5. The average index difference between entrants and graduates is negatively related to the average index of entrants.

3 Defining Students' Academic Indices and University-Major PPI

3.1 Students' Academic Indices

We begin by constructing academic indices for individual students. The first step is to regress graduation outcomes on students' academic qualifications prior to college entry:

$$Y_{ijmt}^{G} = \beta_0 + (ACTM_i * \mathbf{I_i^G})\boldsymbol{\beta}_\mathbf{1}^\mathbf{G} + (ACTR_i * \mathbf{I_i^G})\boldsymbol{\beta}_\mathbf{2}^\mathbf{G} + (CR_i * \mathbf{I_i^G})\boldsymbol{\beta}_\mathbf{3}^\mathbf{G} + \gamma_t + \theta_{jm} + \varepsilon_{ijmt} \quad (1)$$

In equation (1), Y_{ijmt}^{G} is an indicator for whether student *i* in year-cohort *t*, who entered the system in the university-by-major cell defined by university *j* and major *m*, completed a degree in any field within eight years of entry. The variables $ACTM_i$ and $ACTR_i$ are the student's math and reading ACT scores, and CR_i is the student's class percentile rank in high school. The variable vector $\mathbf{I_i^G}$ is a vector of binary indicators for major groupings, denoted by the superscript *G*, with the entry set to one for the major-grouping that encompasses student *i*'s specific major and the other entries set to zero. This feature of the model permits some flexibility in the returns to pre-entry qualifications across majors and is described in more detail in the next paragraph. γ_t is a cohort fixed effect, θ_{jm} a fixed effect for the university-by-major cell, and ε_{ijmt} is an error term, which we specify as having a Type I extreme value distribution implying that the probability of graduation follows a logit. This model is similar to the one developed by Arcidiacono and Koedel (2014).

The superscript *G* indicates the major-group, operationalized through $\mathbf{I_i^G}$, which gives the model flexibility in allowing the qualification measures (ACT math and reading scores and the class rank) to differentially predict success for students who enter in different fields. A model with complete flexibility would allow differential returns across all university-by-major cells, but the parameter space would be large and statistical power limited. Our compromise is to group majors at entry into seven broad categories indexed by *G*: Biological, Mathematical, Physical, & Health Sciences; Business; Education; Engineering and Computer Science; Liberal Arts; Social Science; and Undecided.[10] The model as specified permits major-group specific returns to the three qualification measures, which improves model performance relative to a model that does not allow for parameter heterogeneity by major-group *G* (results omitted for brevity).[11] That said, as we show in the appendix (Appendix Tables A.4 and A.5), a sparse version of the model that does not allow for this type of heterogeneity yields substantively similar conclusions in our analysis of high schools.

We use the output from equation (1), and in particular our estimates of $\boldsymbol{\beta}_1^{\mathbf{G}} - \boldsymbol{\beta}_3^{\mathbf{G}}$, to construct an academic index of pre-entry qualifications, *AI*, for each student as follows:

$$AI_i^G = (ACTM_i * \mathbf{I_i^G})\hat{\boldsymbol{\beta}}_1^{\mathbf{G}} + (ACTR_i * \mathbf{I_i^G})\hat{\boldsymbol{\beta}}_2^{\mathbf{G}} + (CR_i * \mathbf{I_i^G})\hat{\boldsymbol{\beta}}_3^{\mathbf{G}} \qquad (2)$$

The index is a weighted average of a student's pre-entry academic qualifications, where the weights are major-group specific and empirically derived from the graduation model in equation (1) so that the pre-entry qualifications that best predict success (as measured by graduation) are given more weight. Put another way, a higher value for the academic index means that a student's pre-entry qualifications make her more likely to succeed among students who enter the university system in the same major group, all else equal. A critical aspect of the index is that by the inclusion of γ_t and θ_{jm} in equation (1), we ensure

[10] These groupings are exhaustive; that is, each unique major in the system is assigned to one of the groups.
[11] Model performance is improved in the sense that graduation outcomes are predicted more accurately. The heterogeneity afforded by our specification is similar in spirit to heterogeneity in the model used by Arcidiacono and Koedel (2014).

that the identifying variation for the weighting parameters ($\boldsymbol{\beta}_1^{\mathbf{G}} - \boldsymbol{\beta}_3^{\mathbf{G}}$) comes from within university-by-major cells and cohorts.[12]

Table 2 shows results from the estimation of equation (1) – in particular, the coefficient values used to construct the academic index in equation (2) – to provide a sense of the relative importance of students' pre-entry academic qualifications in shaping the index. Focusing on the estimates from our preferred specification in column 1, a general takeaway is a student's class percentile rank is the strongest predictor of graduation conditional on the entering cell. For example, a one standard deviation change in the class rank corresponds to a change in the index of 0.56 to 0.73 depending on major group, whereas standard deviation changes in ACT math or reading scores correspond to index changes on the order of about 0.01 to 0.20. The point estimates on the ACT reading score are generally negative in column 1, but this is because we also condition on high school class rank – ACT reading scores positively predict graduation independently, as shown in the later columns of the table.[13]

The model in column 2 excludes the class rank, which means that no locally-normed information is used to construct the index. While this is not our preferred approach because class rank is the strongest predictor of college success in our data (also see Bowen, Chingos, and McPherson, 2009; Fletcher and Tienda, 2010; Rothstein, 2004), the sparser index formulation can be useful for interpretation. For example, a key finding below is that students from lower-SES high schools enroll in lower-PPI university-by-major cells conditional on their own index values. One explanation is that a high class rank at a low-SES high school is a weaker indicator of academic preparation, which we explore below using the sparser academic index shown in column 2.

[12] We exclude explicit measures of high school quality (high school fixed effects) from the graduation model and index. This allows for a more straightforward examination of the explanatory power of high schools over student placements below.
[13] To provide additional intuition about the index, Appendix Table A.3 replicates Table 2 using a sparse version of the index that does not allow for major-group interactions. With the sparse index it is easier to see the relative weights of the different index components, which without the interactions are interpretable as sample averages across all major groups.

3.2 Preparation and Persistence Indices for University-by-Major Cells

The PPI for each university-by-major cell is based on the academic index values of individuals who complete a degree in that cell, regardless of the entering cell. Therefore, variation in PPI across cells arises from differences in initial selection (which can be driven by students' own choices and the behavior of admissions officials), student persistence within cells, and cross-cell student transfers. We start by taking the average academic index among degree completers in cell *jm*:

$$Q_{jm} = \frac{1}{N_{jm}} \sum_{i=1}^{N_{jm}} AI_i \tag{3}$$

where N_{jm} is the number of individuals who complete a degree in the cell defined by university *j* and major *m*.[14] We then define δ_{jm}, an empirical Bayes estimate for cell *jm*, as follows:

$$\delta_{jm} = \alpha_{jm} * Q_{jm} + (1 - \alpha_{jm}) * Q_j \tag{4}$$

In equation (4), Q_j for university *j* is defined analogously to Q_{jm} as shown in equation (3), but at the university level, and is treated as deterministic. The parameter α_{jm}, with $0 < \alpha_{jm} < 1$, shrinks the overall PPI estimate for cell *jm* toward the university mean (i.e., the prior). The degree of shrinkage depends on the precision with which Q_{jm} is measured, with more-precisely measured values corresponding to higher values of α_{jm}. The formula we use for α_{jm} is:

$$\alpha_{jm} = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\lambda}_{jm}} \tag{5}$$

In equation (5), $\hat{\sigma}^2$ is an estimate of the true variance of Q across university-by-major cells, net of sampling variance, and $\hat{\lambda}_{jm}$ is an estimate of the estimation-error variance of Q_{jm}.

[14] We drop the G superscript on *AI* in equation (3), and in all subsequent references, for notational brevity.

To estimate the parameters used in equation (5) we draw on the recent literature on teacher quality (Koedel, Mihaly and Rockoff, 2015). Briefly, we first estimate the following supplementary regression using degree completers in our analytic sample:

$$AI_{ijm} = \pi_0 + \mathbf{D_{jm}}\boldsymbol{\pi_1} + e_{ijm} \qquad (6)$$

where AI_{ijm} is the academic index for individual *i* who completes a degree in cell *jm*, $\mathbf{D_{jm}}$ is a row vector of indicators for cells, $\boldsymbol{\pi_1}$ is the corresponding column vector, and e_{ijm} is the error term. Intuitively, the variance of π_{1jm} – where π_{1jm} is an entry in the vector $\boldsymbol{\pi_1}$ – gives the variance of Q across cells. Put another way, if the variance of π_{1jm} was zero it would imply no sorting. This variance can be estimated by the variance of $\hat{\pi}_{1jm}$, but the estimate overstates the true variance because it includes sampling variance. Therefore, we adjust the raw variance to obtain an estimate of the true variance of Q – $\hat{\sigma}^2$ in equation (5) – by netting out the sampling variance using the procedure outlined in Koedel (2009).[15] We estimate λ_{jm} from equation (5) as the square of the standard error of $\hat{\pi}_{1jm}$ from equation (6).

This shrinkage procedure is useful analytically because in its absence, variation in cell size across the system generates differential sampling variance in Q_{jm}. For our analysis of high schools the benefit is in the form of improved estimator precision because cell-level PPI is used as the dependent variable. Correspondingly, the findings from our analysis of high schools are qualitatively unaffected if we do not use the shrunken measures, δ_{jm}. However, we also estimate a specification below that maps initial-cell PPI to final-cell PPI among degree completers; for this specification, where measures of PPI are on both the

[15] Koedel's procedure is similar to related procedures found in other studies such as Aaronson, Barrow and Sander (2007), but is better suited to handle situations where there is larger sample-size variance across units (in this case a unit is a university-by-major cell). The adjustment is as follows: $\text{var}(\pi_{jm}) = \text{var}(\hat{\pi}_{jm}) - (\text{var}(\hat{\pi}_{jm})/A)$, where A is a scaled Wald statistic from the test for statistical significance of the full vector of parameters $\boldsymbol{\pi_1}$. See Koedel (2009) for more information; also see Mas and Moretti (2009), who make this adjustment in a technically similar but substantively different context.

left- and right-hand side of the equation, the use of the shrunken measures is necessary to mitigate attenuation bias (Chetty, Friedman and Rockoff, 2014; Jacob and Lefgren, 2008).

Appealing aspects of PPI are its objectivity and flexibility. In terms of objectivity, as noted in the introduction PPI is not influenced by subjective assessments of colleges or majors, either within or across universities, as it depends entirely on the pre-entry academic qualifications of graduates. In terms of flexibility, Figure 2 documents the overlap in cell-level PPI across universities (by selectivity) and between traditionally-classified STEM and non-STEM majors. While the distribution means are ordered as expected, there is considerable distributional overlap along both dimensions. We list the ten highest- and lowest-PPI cells in the Missouri system in Appendix Table A.2 for illustrative purposes.[16]

While these advantages of PPI are useful for our study, we also acknowledge limitations of PPI. Most notably, it should not be interpreted as a comprehensive measure of cell "quality" because there is not a value-added component of PPI. PPI will also be sensitive to the choice of the dependent variable in equation (1), which drives the *AI* weighting parameters ($\boldsymbol{\beta}_1^G - \boldsymbol{\beta}_3^G$). We examine the sensitivity of our findings in this regard by also using 4-year and 6-year graduation outcomes, and first year college GPAs (we restrict our attention to first-year GPAs to avoid the potential confounding issue of differential persistence on cumulative GPA outcomes), in place of 8-year graduation outcomes in equation (1). These changes modify the weights per the procedure described thus far, but none of our findings are substantively affected by using the alternative measures of college success in place of 8-year graduation outcomes (see below for details).[17]

[16] There are some system cells in which students enter but none graduate – the most prominent example includes students who initially enroll as an undecided major. We cannot construct PPI measures using our base methodology for these cells because our measures depend on completers. As an alternative, we construct analogous measures of entry-cell PPI that are a weighted average of final-cell PPI among completers, who by construction must have switched to a different cell. This is an imperfect but functional solution to permit the inclusion of these individuals in our sample. Below we examine the robustness of our findings to dropping students who enter these cells and we obtain similar results.

[17] Approximately 9% of students do not have first year GPAs, but we construct an AI for these students using the parameters estimated by equation (2). An interesting extension of the approach would be to use post-college earnings as the outcome in equation (1), but we do not have access to wage data to pursue this line of inquiry here.

4 Variation in University-by-Major PPI and Student Sorting

A basic variance decomposition of cell-level PPI indicates that 62 percent of the variance occurs across universities and 38 percent occurs within. While this split affirms the literature's focus on the importance of institutional sorting (Arcidiacono and Lovenheim, 2016; Dillon and Smith, 2017; Hoxby and Turner, 2014; Smith, Pender, and Howell, 2013), it also highlights the presence of substantial variability in major PPI within institutions.

In addition to the decomposition, we also use measures of academic alignment between individual students and their initial university-by-major cells to contextualize system sorting. To do so, we first define academic alignment for student *i* who enters cell *jm* as $M_{i,jm} = AI_i - \delta_{jm}$. We compare observed alignment based on actual student sorting to alignment under two types of counterfactual sorting conditions: (1) random assignment of students to system cells; and (2) perfect sorting of students to system cells (where we assign the highest-*AI* students to the cells with the highest values of δ_{jm}). For each set of counterfactual conditions, we consider two scenarios: (a) a "global" scenario in which the counterfactual sorting occurs across and within universities; and (b) a "local" scenario where the counterfactual sorting is conditional on the initial university. For example, with global random assignment, we randomly assign students to majors and universities; whereas with local random assignment, we randomly assign students to majors holding the entering university fixed. The variance of the alignment measure, $M_{i,jm}$, will be minimized in the global perfect-sorting case because students' own academic indices will align most closely with the hypothetical entering university and major.[18] The variance will be at its practical maximum with global random assignment. These comparisons provide context for observed sorting.

Table 3 reports the results. The top row shows the variance of $M_{i,jm}$ based on students' actual university-by-major placements. Subsequent rows report the variance under the four counterfactuals. The

[18] This minimization is subject to the pre-existing structure of the system, and in particular the size of system cells, which we hold fixed for this descriptive analysis.

observed variance of $M_{i,jm}$, 0.50, falls comfortably between the two global counterfactual bounds of 0.22 (perfect sorting) and 0.93 (random sorting).

The counterfactual scenarios provide useful insight into the potential for cross-university and within-university sorting to affect alignment. For example, the within-university, perfect sorting condition minimizes within-university misalignment (last row of Table 3). The variance of $M_{i,jm}$ in this scenario is 0.30, which is close to the global perfect-sorting condition (0.22); certainly much closer than the observed sorting condition (0.50). The implication is that resorting students to majors with closer academic alignment, without any switching across universities, would increase alignment nearly as much as resorting students across the entire system. This does not diminish the importance of college placements in studying postsecondary sorting, but it does motivate the importance of also studying sorting within universities.

5 The Role of High Schools in Student Sorting

Having defined each student's own preparation index and the PPI of the entering university-by-major cell, we examine the explanatory power of high schools over student placements into colleges and majors conditional on each student's own academic preparation. We start with the following linear regression model:

$$\delta_{jm,is} = \gamma_0 + AI_i\gamma_1 + \mathbf{HS_{is}}\boldsymbol{\gamma_2} + u_{jm,is} \tag{7}$$

In equation (7), the PPI of university-by-major cell *jm* into which student *i* from high school *s* enters, $\delta_{jm,is}$, is a function of the student's own academic index, AI_i, and the high school attended, where $\mathbf{HS_{is}}$ is a vector of indicator variables in which the student's own high school indicator is set to one and all others are set to zero. We do not allow a student's own academic index to contribute to $\delta_{jm,is}$ to prevent spurious correlations. Thus, if a student starts and completes a degree in cell *jm*, her own academic index is jack-knifed out of the calculation of $\delta_{jm,is}$. The parameter γ_1 is identified using within high-school variation in AI_i to estimate the empirical relationship between a student's own academic preparation and the PPI of the initial cell. Conditional on this relationship, the vector of high school fixed effects, $\boldsymbol{\gamma_2}$, captures systematic differences in the PPI of placements across high schools. $u_{jm,is}$ is the residual in the regression. We estimate standard errors using a 2-way clustering structure to account for dependence in the data within university-by-major cells and high schools following Petersen (2009; also see Cameron and Miller, 2015).

The model in equation (7) can be adjusted to examine the extent to which high schools explain differences in the PPI of student placements across majors *within universities* as follows:

$$(\delta_{jm} - \delta_j)_{is} = \theta_0 + AI_i\theta_1 + \mathbf{HS_{is}}\boldsymbol{\theta}_2 + e_{jm,is} \tag{8}$$

The only change in equation (8) is the dependent variable is measured relative to overall university PPI, where universities are subscripted by *j*. Our measures of university PPI are constructed analogously to our measures of university-by-major PPI per the description in Section 3.[19]

Next we examine whether characteristics of high schools systematically explain the PPI of student placements. Following on previous research showing that students from disadvantaged backgrounds tend to enroll in universities where their own academic preparation exceeds that of their peers, we are particularly interested in the degree to which measures of socioeconomic disadvantage at the high school level predict placement PPI. To investigate this question we estimate the following analogs to equations (7) and (8):

$$\delta_{jm,is} = \rho_0 + AI_i\rho_1 + \mathbf{Z_{is}}\boldsymbol{\rho_2} + \xi_{jm,is} \tag{9}$$

$$(\delta_{jm} - \delta_j)_{is} = \psi_0 + AI_i\psi_1 + \mathbf{Z_{is}}\boldsymbol{\psi_2} + \zeta_{jm,is} \tag{10}$$

These equations substitute high school and local-area characteristics, in the Z-vector, for the high school indicators in equations (7) and (8). The measures of socioeconomic disadvantage that we include are the share of the student body eligible for free or reduced price lunch (FRL) and the share of individuals age-25 and older with less than a bachelor's degree in the high school's zip code. We also include the share of the student body that identifies as a minority race or ethnicity. In addition to these focal high-school characteristics, we condition on basic characteristics of high schools including urbanicity (schools are

[19] In fact, because we treat university PPI as deterministic per Section 3, $\delta_j = Q_j$

divided into five groups: urban, suburban, town, rural and missing) and school size (enrollment), along with a vector of three geography-based variables meant to capture the geographic placement of each high school with respect to the university system.[20] The three geography-based controls are: distance to the nearest university, university-level PPI of the nearest university, and the interaction between the two.

For ease of interpretation, we normalize the dependent variables and high-school characteristics throughout to have a mean of zero and a variance of one.[21] In our preferred specifications as shown in equations (7)-(10), we also normalize the academic index for individuals and enter it into the models linearly. In Appendix Table A.6 we show that our findings are qualitatively unaffected if we use a more flexible modeling approach where we divide students into twenty equal-sized bins based on their own index values and condition on bin assignment instead.

6 Results

We assess the general importance of high schools and their surrounding areas in explaining students' initial placements conditional on their own academic indices using equations (7) and (8). Table 4 reports the overall R-squared and partial R-squared attributable to the vector of high school indicators for each model. The table shows that high schools explain 10.4 percent of the variance in university-by-major PPI overall. However, they explain just 1.7 percent of the within-university variance, implying that their explanatory power is primarily over university placements.

We compare the explanatory power of the high school indicators reported in Table 4 to the explanatory power of observed high-school characteristics to determine how much of the predictive influence of high schools is explained by our vector of observables. We obtain the explanatory power of high school characteristics similarly to the high school fixed effects, using the partial R-squared – i.e., we

[20] School-level observable characteristics are taken from the Common Core of Data (CCD) and the local-area characteristics are from the year-2000 U.S. Census.

[21] More precisely, the dependent variables are normalized so that a one-unit change represents a one-standard-deviation change in the true distribution of PPI. In practice, the normalized dependent variables have a standard deviation of less than one because they are normalized by the un-shrunken standard deviations. This facilitates the interpretation of a one-unit change in PPI as corresponding to a one standard deviation change in the true (rather than empirical) distribution (see also Chetty, Friedman, Rockoff, 2014; Jacob and Lefgren, 2008).

start with a model that just includes students' own AI values, then add the high school characteristics and capture the increase in the R-squared. The high-school and local-area SES variables, along with school size and the urbanicity indicators, explain 5.8 percent of the variance in PPI overall, or roughly 56 percent of the variance explained by high schools in total as shown in Table 4 (5.8/10.4). Adding the vector of geographic controls increases the partial R-squared from 5.8 to 6.9 percent. Thus, overall, we can account for 66 percent of the explanatory power of high schools with the observable characteristics available to us (6.9/10.4). In contrast, high school characteristics account for only a very small fraction of the variance in PPI within universities explained by the high school indicators, which per Table 4 is already minimal. Specifically, the partial R-squared attributable to our full set of high school characteristics, inclusive of the geography variables, in the within-university sorting model is just 0.003; which means that these variables explain just 20 percent of the total variance explained by high schools (0.3/1.7).

Next, in Table 5 we show results from variants of equation (9) where we replace the high school indicators with high school characteristics to document the relationships between student sorting and high-school and local-area SES. We include the minority share and each measure of socioeconomic disadvantage in the model separately and then include them all simultaneously, with and without conditioning on the other non-SES high school controls. In the full specification in the final column of Table 5, one standard deviation increases in the minority share, the percentage of FRL-eligible students, and the share of the local area with less than a bachelor's degree correspond to changes in the PPI of the initial university-major cell of 0.01 (not statistically significant), -0.03, and -0.12 standard deviations, respectively. A general takeaway from Table 5 is that students from more disadvantaged backgrounds sort to lower PPI university-by-major cells conditional on their own academic preparation, which is in line with previous research on undermatch to universities (Turner, 2017).[22]

[22] As noted above, while the high school minority share does not predict cell PPI in the full sample conditional on the other controls, if we exclude students who attend the HBCUs from our sample there is a modest positive relationship between cell PPI and the high school minority share. The implication is that the HBCUs disproportionately lower cell PPI for students from high-minority high schools, which follows from their low PPI-based rankings and relatively high minority enrollment shares.

Next we extend the analysis to look for systematic placements by high school minority share and SES within universities. Table 6 follows the same structure as Table 5, but focuses on within-university placements per equation (10). Consistent with the limited explanatory power of high schools over within-university sorting documented in Table 4, and the limited explanatory power of observed high school characteristics as reported above, the results in Table 6 provide no indication of differences between students from high schools with different characteristics. None of the high-school SES measures are meaningfully associated with placements by PPI within universities, individually or jointly.

As noted briefly above, we also replicate the analysis described thus far using "sparse" versions of equations (1) and (2) that do not allow for heterogeneity in the returns to academic qualifications by major-group (indexed by *G* in the equations). The sparse model is useful for investigating the extent to which match quality between students and majors drives our findings thus far, in that unlike our preferred specification, it does not allow for match quality effects. Results analogous to those shown in Tables 4, 5 and 6, but generated based on the sparse versions of equations (1) and (2), are reported in Appendix Tables A.4 and A.5. They are very similar to our primary findings, indicating that issues related to match quality between students and majors do not drive our findings.

7 Robustness

7.1 The Use of Alternative College Outcomes to Determine the AI Weights

As noted previously, the construction of students' academic indices, and correspondingly cell-level PPI, depends on the outcome measure used in equation (1). It is this outcome measure that determines the weighting parameters for the pre-entry qualification measures, $\boldsymbol{\beta}_1^{\mathbf{Q}} - \boldsymbol{\beta}_3^{\mathbf{Q}}$. The outcome we have used thus far is the 8-year graduation. In this section we consider the sensitivity of our findings to using alternative AI and PPI constructs based on 4-year and 6-year graduation outcomes, and first-year GPAs. For each alternative outcome we begin by re-estimating equation (1) to get new weighting parameters for students' individual academic indices, then go through the entire analytic procedure outlined above.

For brevity we relegate tables with the results to the appendix (see Appendix Tables A.7 and A.8), but none of the findings from our analysis of high schools and their local areas are substantively affected by changing the outcome in equation (1). More specifically, the explanatory power of high schools over system-wide PPI sorting, and sorting within universities, is similar, as are the relationships between observable high school and local-area characteristics and student sorting. We conclude that our results are qualitatively robust to using alternative measures of postsecondary success as the foundation for our analysis.

7.2 Cells without Completers

Next we turn to the issue that approximately one-third of the students in the sample enter into cells in which there are no completers. The predominant example is students who list their initial field of study as "undecided," who account for about one-fifth of our sample, or approximately 13,000 students. There are also another 5,800 students who begin in a cell without any finishers, with the most common reason being that the initial cell is a broad field such as "general engineering." Students who enter into a broad field like "general engineering" do not finish with a general degree. Instead, they either finish in a more specific engineering subfield, such as chemical engineering or mechanical engineering, switch to a completely different discipline, or drop out. In the analysis thus far, we have handled such cells by assigning them a PPI measure that is a weighted average of *finishing cell* PPI across all graduating students who enter. This is a functional solution, but treats these cells differently than other cells (for other cells, only finishers matter regardless of the entering cell as described in Section 3.2).

In Appendix Table A.9 we examine the sensitivity of our findings to dropping all students who enter university-by-major cells with no completers, since we do not have a consistent strategy for constructing measures of cell PPI for these students. For brevity, we replicate our estimates from the full models shown in Tables 5 and 6 only. The results show that our findings are qualitatively unaffected by whether we include these individuals in the analysis.

8 Extensions

8.1 Heterogeneity Among high- and low-AI Students

In this section we briefly ask whether high schools differentially predict sorting between above- and below-median AI students. To answer this question we replicate our primary findings from Tables 4-6 separately for subsamples of students with above- and below-median AI values. The results are reported in Tables 7 and 8.

The tables show that our findings are directionally consistent for above- and below-median students, and the general takeaway that high schools (and their characteristics) explain a substantial fraction of university sorting, but not sorting to majors within universities, is upheld for both student subsamples. An interesting disparity that emerges is that students' own AI values are much stronger predictors of sorting among high-AI students than among low-AI students, both systemwide and within universities. This can be seen clearly in the first row of estimates in Tables 7 and 8. The substantial gap between low- and high-AI students in the correspondence between their own measures of preparation and sorting behaviors suggests very different sorting processes.[23]

8.2 An Alternative Academic Index Excluding Class Rank

The measure of academic preparation that receives the most weight by far in the individual academic index – the high school class percentile rank – is a locally-normed measure. While it is well-established that high school performance is a stronger predictor of college success than entrance exam scores (in addition to our results above, also see Bowen, Chingos, and McPherson, 2009; Fletcher and Tienda, 2010; Rothstein, 2004), the fact that it is locally normed creates some ambiguity in the interpretation of our findings. For example, a reason we might find that students from low-SES high schools enter the system in lower-PPI cells is that conditional on the index, their preparation is lower than that of their high-SES peers. Put differently, it may be that performing at the top of the class at a low-SES high school does

[23] The substantial differences in the coefficients on own-AI when we split the sample partly reflect differential coverage over the support of cell-level PPI for the two student subgroups. Unsurprisingly, high-AI students are more concentrated among high-PPI cells and the reverse is true for low-AI students.

not signify the same level of preparation as performing at the top of the class at a high-SES high school. This possibility is consistent with findings from Black et al. (2015), who show that students in Texas with high class ranks but who attended low-performing high schools have persistently lower grades throughout college than their peers who attended better high schools.

We gain some insight into this issue by using a version of the academic index that does not include the class percentile rank, from column 2 of Table 2. We present the results in Table 9, where we replicate our full procedure and show specifications akin to those in Tables 5 and 6 using the restricted academic index. Again noting the caveat to these results that we sacrifice substantial informational content by excluding information about students' class ranks, in the model examining system-wide placements in columns 1 and 2, we find directionally similar but weaker results to what we show in Table 5 for the income and education SES measures, but the coefficient on the minority is larger and statistically significant. This pattern of results is also apparent when we enter the high-school SES and minority share measures into the models separately (not shown for brevity). In columns 3 and 4, where we replicate the results from Table 6, there is also a moderate shift toward the appearance of less under-placement for students from low-SES high schools. Specifically, whereas with our primary specification there is not a detectable pattern of within-university sorting by high school SES conditional on students' own academic preparation, when we use the restricted index we find that students from lower-SES high schools conditionally enroll in modestly *higher-PPI* majors within universities. In summary, students from low-SES high schools seem less under-placed when we no longer account for class rank.

This shift in results is consistent with the interpretation that our primary estimates in Tables 5 and 6 are driven in part by the fact that highly ranked students from low-SES high schools are not as well prepared as their highly ranked peers from high-SES high schools. Either by their own application and enrollment actions or the actions of university admissions officials, this is reflected in lower-PPI placements conditional on these students' academic indices. This interpretation has significant social meaning: the unequal value of class rank would directly imply that differential opportunities for human capital development during K-12 schooling between students in high- and low-SES high schools explain some of the differences we observe in entering-cell PPI.

8.3 The Mapping Between Initial and Final University-by-Major Cells

Students' initial placements influence their academic experiences and outcomes (e.g., Artz and Welsch, 2014; Carrell, Fullerton, and West, 2009; Porter and Umbach, 2006; St. John et al., 2004). However, there is also a robust literature that connects post-college outcomes to *final* college and major (Arcidiacono, 2004; Carnevale et al., 2016; Eide, Hillmer, and Showalter, 2015; Hamermesh and Donald, 2008; Thomas and Zhang, 2005; Webber, 2016). An obvious question given our focus on initial university-by-major placements is how initial placements translate to final placements.

To answer this question we begin with basic summary statistics. Among students who declared a major when they entered the system and graduated, almost 40 percent finished in the same cell that they entered. Furthermore, nearly 60 percent finished in the same major group (with the same 2-digit CIP code) as the entering major. These numbers suggest initial placements have significant inertia.

To address this question more generally, we estimate the relationship between the PPI of the initial and final cell using a simple, student-level regression of the following form:

$$\delta^F_{jm,is} = \varphi_0 + \delta^I_{jm,is}\varphi_1 + AI_i\,\varphi_2 + \zeta_{jm,is} \tag{11}$$

In equation (11), $\delta^F_{jm,is}$ is the normalized PPI of the final cell and $\delta^I_{jm,is}$ is the normalized PPI of the initial cell.[24] The estimation of equation (11) is restricted to degree completers.

First, Figure 3 plots the unconditional relationship between $\delta^F_{jm,is}$ and $\delta^I_{jm,is}$ among completers. The markers represent the average ending PPI for each bin of beginning PPI, with bin sizes of 0.1 standard deviations. The size of each marker reflects the number of students in the bin. It is visually apparent that the PPI of the initial major is highly predictive of the PPI of the final major, and that this strong relationship

[24] As in the preceding analysis, the normalizations are performed to facilitate interpretations in terms of the real (rather than empirical) distributions of PPI. Because the PPI measures are shrunken, estimates of φ_1 will not be affected by attenuation bias (Jacob and Lefgren, 2008).

holds throughout the distribution of beginning-cell PPI. This is supported formally by results from equation (11), where we estimate φ_1 to be 0.93 with a standard error of 0.05.

The strong link we identify between PPI of the starting and ending cells should not be interpreted causally and it is important not to infer that simply changing initial placements will necessarily change final placements. That said, the link is quite strong, which implies policies that change students' initial placements *and the factors that underlie these placements* can meaningfully change the distribution of university-by-major exit pathways.

9 Conclusion

We use empirical measures that capture dimensions of selectivity and rigor at the university-by-major level to examine the explanatory power of high schools over students' college and major placements. Our measures – which we term "preparation and persistence indices" (PPIs) – are based on students' weighted pre-college academic qualifications, where the weights are determined by a regression of college graduation outcomes. PPI affords us flexibility in examining student sorting within the 4-year public university system in Missouri and it varies substantially both within and across universities.

Our examination of the explanatory power of high schools and their local areas over students' initial university-by-major placements, conditional on students' own academic preparation, yields the insights that they explain (a) a substantial share of the variance in the PPI of university placements, and (b) little of the variance in the PPI of major placements within universities. Corroborating previous research, the socioeconomic status of high schools and their local areas is a clear predictor of the PPI of students' initial university placements, with lower-SES students systematically enrolling at lower-PPI universities conditional on their own academic preparation (Dillon and Smith, 2017; Hoxby and Avery, 2013; Hoxby and Turner, 2014; Smith, Pender, and Howell, 2013). When we examine sorting using PPI measures that exclude locally-normed information about class-rank, the sorting differences by high school SES moderate, which is consistent with the explanation that differential access to K-12 school quality accounts for part of the gap in students' initial college placements.

The mechanisms that account for the differential explanatory power of high schools over university sorting, versus within-university sorting to majors, merit more attention in future research. Delving into these mechanisms is largely outside of the scope of the current paper, although we do show that high school characteristics also explain much more of the variation in university sorting – it is not just unobserved factors associated with high schools that account for the difference. An intuitive hypothesis is that geography impacts university sorting but not major sorting within universities. Using somewhat rough controls that situate each high school within the context of the higher education system geographically, we see some support for this hypothesis: these controls account for about 1.1 percent of the variance in systemwide PPI placements, but a miniscule 0.07 percent of the variance in PPI placements within universities. That said, this is clearly not the whole story, as even after accounting for this difference high schools continue to explain much less of the variance in within-university sorting.

Our findings have several important implications for research and policy. First, they point toward the value of interventions that inform students of the educational options for which they are academically qualified, which can better align students from low-SES high schools with universities (Hoxby and Turner, 2014). Heterogeneity in student preferences ensures that under- and over-placements to universities will occur, especially since non-academic factors also play an important role in determining the college match (Bond et al., forthcoming). However, the systematic relationship between under-placement and student SES we document is disconcerting in light of evidence that more-selective institutions, as measured by the academic qualifications of entering students, improve educational outcomes (Arcidiacono and Koedel, 2014; Cohodes and Goodman, 2014; Hoekstra, 2009; Melguizo, 2010).[25] Moreover, even if some of the disparate sorting behavior between seemingly similarly-qualified students from high- and low-SES high

[25] Much of the research on potentially harmful effects of students attending institutions for which they may not have adequate observed preparation relates to affirmative action policies, for which there is limited evidence of an academic penalty, per Arcidiacono & Lovenheim (2016). These authors generally report a positive return to college quality for both graduation likelihood and labor market outcomes, though less-prepared students may end up in relatively less rigorous majors than their peers (e.g., see Arcidiacono, Aucejo and Spenner, 2012). Related to this issue, Dillon and Smith (2017) show that the preferences of more informed students (and their families) imply that they believe the benefits of college quality more than offset any possible costs associated with over-placement.

schools is driven by true gaps in student preparation owing to unequal opportunities during high school (per Table 9), the greater efficacy of more selective institutions will still likely benefit lower-SES students.

Second, despite high schools offering little explanatory power over within-university sorting, we document substantial within-university variation in PPI between majors. Majors can affect learning and influence students' academic environments, including interactions with faculty and the development of peer groups (e.g., Artz and Welsch, 2014; Carrell, Fullerton, and West, 2009; Porter and Umbach, 2006; St. John et al., 2004). But little is known about the practical importance of quality differences across majors in terms of affecting student outcomes, or about the malleability of student allocations to departments within universities should reallocations be desirable. Our findings at least raise the possibility that, like with the aforementioned recent literature on college selectivity, postsecondary educational production could be improved by more purposeful allocations of students to majors within universities. Said another way, students across the ability distribution may benefit from placements in high quality majors; future research probing the significance of within-university variability in major quality and student sorting can shed light on this issue.

Finally, we show that initial university-by-major PPI is a strong predictor of final university-by-major PPI among degree completers. This is driven in part by cell persistence, but it is also the case that cell changes tend to be PPI-aligned. An implication is that a pressure point for policy interventions that aim to affect the skill distribution of the workforce through human capital development in college occurs prior to college entry.

References

Aaronson, Daniel, Lisa Barrow and William Sander. 2007. Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25(1), 95-135.

Arcidiacono, Peter. 2004. Ability Sorting and the Returns to College Major. *Journal of Econometrics* 121(1–2), 343-375.

Arcidiacono, Peter, Esteban Aucejo, and V. Joseph Hotz. University Differences in the Graduation of Minorities in STEM Fields: Evidence from California. *American Economic Review*, 106(3), 525-562.

Arcidiacono, Peter, Esteban Aucejo, and Ken Spenner. 2012. What Happens After Enrollment? An Analysis of the Time Path of Racial Differences in GPA and Major Choice. *IZA Journal of Labor Economics*, 1(5).

Arcidiacono, Peter and Cory Koedel. 2014. Race and College Success: Evidence from Missouri. *American Economic Journal: Applied Economics*, 6(3), 20-57.

Arcidiacono, Peter and Michael Lovenheim. 2016. Affirmative Action and the Quality-Fit Tradeoff. *Journal of Economic Literature*, 54(1), 3-51

Artz, Benjamin, and David M. Welsch. 2014. The Effect of Peer and Professor Gender on College Student Performance. *Southern Economic Journal*: 80(3), 816–838.

Betts, Julian R., and Darlene Morrell. 1999. The Determinants of Undergraduate Grade Point Average: The Relative Importance of Family Background, High School Resources, and Peer Group Effects. *Journal of Human Resources*, 34(2), 268-293.

Black, Sandra, Jane Lincove, Jennifer Cullinane, and Rachel Veron. 2015. Can You Leave High School Behind? *Economics of Education Review* 46, 52-63.

Bond, Timothy, George Bulman, Xiaoxiao Li, and Jonathan Smith (forthcoming). Updating Human Capital Decisions: Evidence from SAT Score Shocks and College Applications. *Journal of Labor Economics*.

Bowen, William G., Mathew M. Chingos and Michael S. McPherson. 2009. *Crossing the Finish Line*. Princeton, New Jersey: Princeton University Press.

Cameron, A. Colin and Douglas L. Miller. 2015. A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2), 317-372.

Cameron, Stephen V. and James J. Heckman. 2001. The Dynamics of Educational Attainment for Black Hispanic, And White Males. *Journal of Political Economy* 109(3), 455-499.

Carnevale, Anthony P., Megan L. Fasules, Andrea Porter and Jennifer Landis-Santos. 2016. African Americans: College Majors and Earnings. Policy Report. Center on Education and the Workforce: Georgetown University.

Carrell, Scott, Richard Fullerton, and James West. 2009. Does Your Cohort Matter? Measuring Peer Effects in College Achievement. *Journal of Labor Economics*, 27(3), 439-464.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2014. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review* 104(9), 2633-79.

Cohodes, Sarah and Joshua Goodman. 2014. Merit Aid, College Quality and College Completion: Massachusetts' Adams Scholarship as an In-Kind Subsidy. *American Economic Journal: Applied Economics* 6(4), 251-85.

Committee on Prospering in the Global Economy of the 21st Century. 2007. *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future*. Washington DC: The National Academies Press.

Deming, David. 2015. The Growing Importance of Social Skills in the Labor Market. NBER Working Paper No. 21473.

Dillon, Eleanor and Jeff Smith. 2017. Determinants of the Match between Student Ability and College Quality. *Journal of Labor Economics* 35(1), 45-66.

Eide, Eric R., Michael J. Hilmer, and Mark H. Showalter. 2016. Is it Where You Go or What You Study? The Relative Influence of College Selectivity and College Major on Earnings. *Contemporary Economic Policy*, 34, 37–46.

Fletcher, Jason. 2013. Social Interactions and College Enrollment: Evidence from the National Education Longitudinal Study. *Contemporary Economic Policy*, 31, 762–778.

Fletcher, Jason, and Adalbert Mayer. 2014. Tracing the Effects of Guaranteed Admission through the College Process: Evidence from a Policy Discontinuity in the Texas 10% Plan. *Contemporary Economic Policy*, 32, 169-186.

Fletcher, Jason and Marta Tienda. 2010. Race and Ethnic Differences in College Achievement: Does High School Attended Matter? *The ANNALS of the American Academy of Political and Social Science* 627: 144-166.

Hamermesh, Daniel S. and Stephen G. Donald, The Effect of College Curriculum on Earnings: An Affinity Identifier for Non-ignorable Non-response Bias. *Journal of Econometrics* 144(2), 479-491.

Hoekstra, Mark. 2009. The Effect of Attending the Flagship State University on Earnings: A Discontinuity-Based Approach. *Review of Economics and Statistics* 91(4), 717-724.

Hoxby, Caroline and Christopher Avery, 2013. The Missing "One-Offs": The Hidden Supply of High-Achieving, Low-Income Students. *Brookings Papers on Economic Activity* 46(1), 1-65.

Hoxby, Caroline and Sarah Turner. 2014. Expanding College Opportunities for High-Achieving, Low Income Students. SIEPR Discussion Paper No. 12-014.

Hurwitz, Michael, Preeya P. Mbekeani, Margaret M. Nipson, and Lindsay C. Page. 2017. Surprising Ripple Effects: How Changing the SAT Score-Sending Policy for Low-Income Students Impacts College Access and Success. *Educational Evaluation and Policy Analysis* 39(1), 77-103.

Jacob, Brian and Lars Lefgren. 2008. Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics* 26(1), 101-136.

Koedel, Cory. 2009. An Empirical Analysis of Teacher Spillover Effects in Secondary School. *Economics of Education Review*, 28(6), 682–692.

Koedel, Cory, Kata Mihaly and Jonah E. Rockoff. 2015. Value-Added Modeling: Review. *Economics of Education Review*, 47, 180-195.

Long, Mark C., Patrice Iatarola, and Dylan Conger. 2009. Explaining Gaps in Readiness for College-Level Math: The Role of High School Courses. *Education Finance and Policy*, 4(1), 1-33.

Mas, Alexandre, and Enrico Moretti. 2009. Peers at Work. *American Economic Review*, 99(1), 112-145.

Melguizo, Tatiana. 2010. Are Students of Color More Likely to Graduate From College if They Attend More Selective Institutions? Evidence From a Cohort of Recipients and Nonrecipients of the Gates Millennium Scholarship Program. *Educational Evaluation and Policy Analysis*, 32(2), 230-248.

Petersen, Mitchell. 2009. Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *Review of Financial Studies* 22(1): 435-480

Porter, Stephen and Paul Umbach. 2006. College Major Choice: An Analysis of Person–Environment Fit. *Research in Higher Education*, 47(4), 429-449.

Rothstein, Jesse M. 2004. College performance predictions and the SAT. *Journal of Econometrics* 121(1-2): 297-317.

Smith, Jonathan, Matea Pender, and Jessica Howell. 2013. The Full Extent of Student-College Academic Undermatch. *Economics of Education Review, 32*.

St. John, Edward, Shouping Hu, Ada Simmons, Deborah Carter, and Jeff Weber. 2004. What Difference Does a Major Make? The Influence of College Major Field on Persistence by African American and White students. *Research in Higher Education*, 45(3), 209–232.

Stinebrickner, Ralph and Todd R. Stinebrickner. 2014. A Major in Science? Initial Beliefs and Final Outcomes for College Major and Dropout. *Review of Economic Studies* 81(1), 426-472.

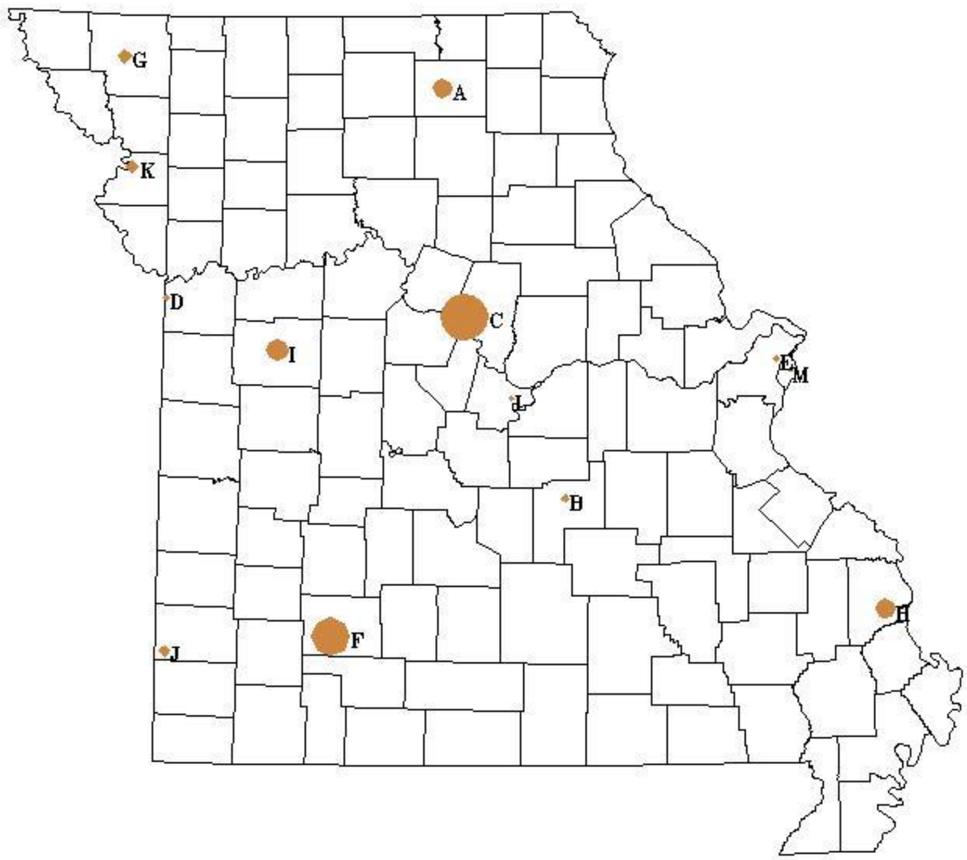
Thomas, Scott and Liang Zhang. 2005. Post-Baccalaureate Wage Growth within Four Years of Graduation: The Effects of College Quality and College Major. *Research in Higher Education*, 46(4), 437-459.

Turner, Sarah. 2017. Education Markets: Forward-Looking Policy Options. Hutchins Center Working Paper #27. Washington, DC: Brookings.

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

Wiswall, Matthew and Basit Zafar. 2015. Determinants of College Major Choice: Identification using an Information Experiment. *Review of Economic Studies*, *82(2)*, 791-824.

Figure 1: Geographic Distribution of 4-year Public Universities in Missouri



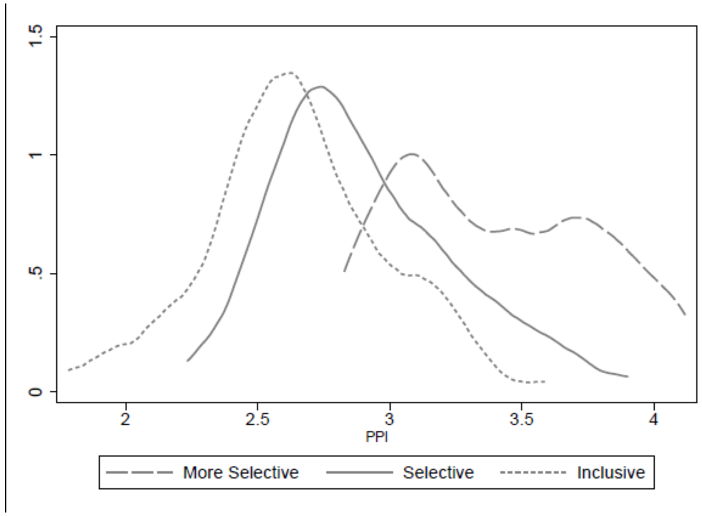
Legend

A: Truman State University
B: Missouri Science and Technology (UM-Rolla)
C: UM-Columbia
D: UM-Kansas City
E: UM-St. Louis
F: Missouri State University
G: Northwest Missouri State University
H: Southeast Missouri State University
I: University of Central Missouri
J: Missouri Southern State University
K: Western Missouri State University
L: Lincoln University
M: Harris Stowe State University

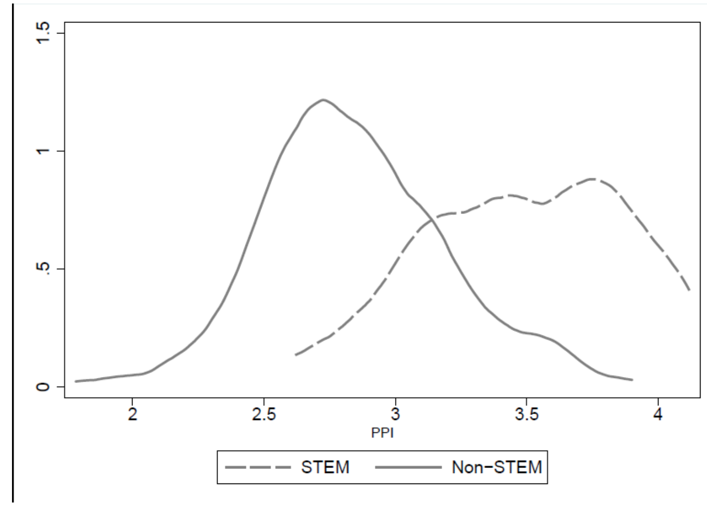
Note: Circle sizes correspond to enrollment shares from the analytic sample.

Figure 2: Distributions of PPI by University Selectivity and Major Category

Panel A: University Selectivity Group

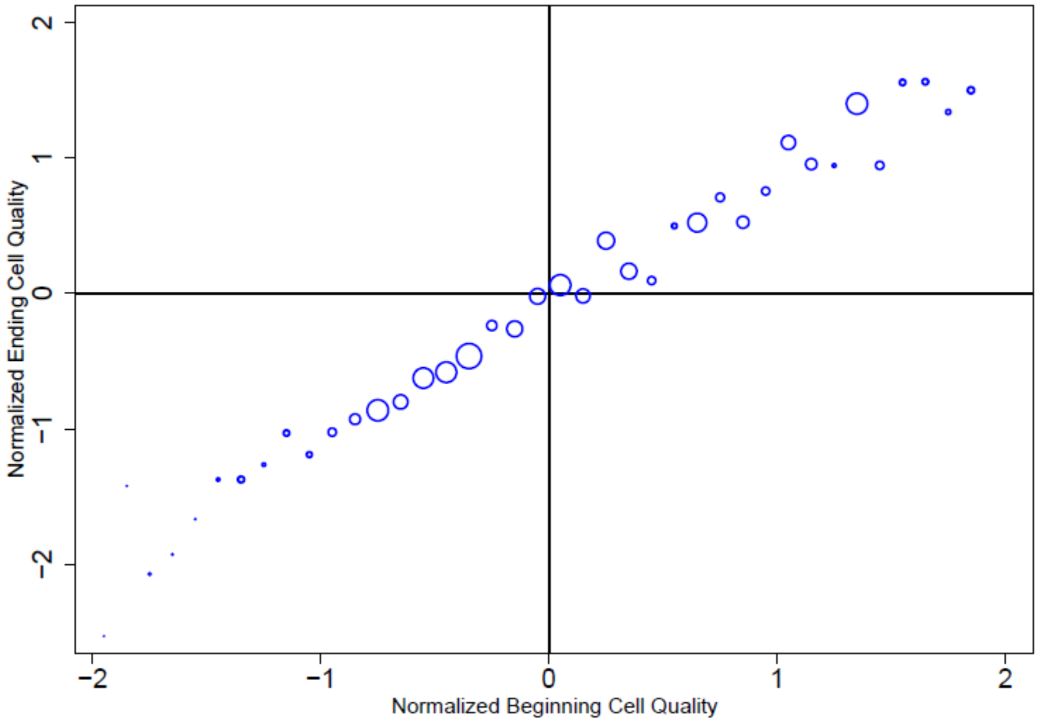


Panel B: STEM and non-STEM Majors



Notes: Panel A shows kernel density plots of PPI by university-selectivity group. The “more selective” institutions include: Missouri Science and Technology (UM-Rolla); Truman State University; University of Missouri-Columbia. “Selective” institutions include: University of Missouri-Kansas City; University of Missouri-St. Louis; Missouri State University; Northwest Missouri State University; University of Central Missouri; and Southeast Missouri State University. “Inclusive” institutions include: Missouri Southern State University; Western Missouri State University; Lincoln University; and Harris Stowe State University. Panel B plots kernel densities of PPI for traditionally defined STEM and non-STEM fields. STEM fields include (2-digit CIP codes in parentheses): Computer and Information Sciences (11); Engineering (15); Biological and Biomedical Sciences (26); Mathematics and Statistics (27); and Physical Sciences (40). The overlap displayed in both graphs is substantively unaffected by reasonable adjustments to the university and major groupings.

Figure 3: Relationship between the PPI of the Final and Initial University-Major Cell



Notes: Graph depicts the relationship between normalized ending university-major cell PPI (on the y-axis) and normalized beginning cell PPI (on the x-axis). Markers are the average ending PPI for the values of beginning PPI, with beginning PPI grouped into bins of 0.1 standard deviations. The size of each marker reflects the number of students in the bin. This chart only includes students who finish.

Table 1. University Descriptive Statistics for Analytic Sample.

University	Average Academic Index Of Entrants	Standard Dev. Academic Index of Entrants	Entry Share	Graduation Rate	Average Academic Index Of Graduates	Standard Dev. Academic Index of Graduates
Overall	2.77	0.73	1.00	0.62	2.97	0.63
Missouri Science and Technology (UM-Rolla)	3.79	0.58	0.04	0.72	3.89	0.54
Univ of Missouri-Columbia	3.32	0.71	0.22	0.75	3.39	0.69
Univ of Missouri -Kansas City	3.22	0.80	0.04	0.55	3.28	0.81
Truman State Univ	3.16	0.57	0.08	0.78	3.21	0.56
Univ of Missouri -St. Louis	2.86	0.74	0.03	0.50	2.91	0.76
Missouri State Univ	2.65	0.75	0.19	0.59	2.83	0.71
University of Central Missouri	2.64	0.76	0.10	0.60	2.82	0.72
Northwest Missouri State Univ	2.61	0.78	0.07	0.64	2.78	0.74
Missouri Southern State Univ	2.45	0.85	0.05	0.44	2.75	0.77
Southeast Missouri State Univ	2.43	0.80	0.09	0.58	2.63	0.76
Western Missouri State Univ	2.22	0.86	0.07	0.41	2.65	0.78
Lincoln Univ	2.06	0.94	0.02	0.39	2.49	0.88
Harris Stowe State Univ	1.91	1.08	0.00	0.30	2.02	1.12

Notes: The analytic sample includes full-time, resident, non-transfer students who entered the system between 1996 and 2001 as college freshman from public high schools. It omits students whose high school of attendance, class rank, and/or ACT scores are unavailable (combined data loss ≈ 6 percent). The enrollment shares presented in this table are broadly reflective of the relative sizes of the public universities in Missouri, but can differ from total enrollment shares because we exclude transfer students from community colleges as well as part-time students, and these students are not evenly distributed across the system.

Table 2. Index Parameters from Primary and Alternative Specifications for the Index.

	Index parameters				
	(1)	(2)	(3)	(4)	(5)
High School Class Rank Interacted with Major Group:					
Biological, Mathematical, Physical, & Health Sciences	3.410 (0.152)		3.725 (0.139)		
Liberal Arts	3.288 (0.153)		3.385 (0.141)		
Engineering and Computer Science	3.474 (0.170)		3.715 (0.159)		
Education	3.351 (0.176)		3.450 (0.159)		
Social Sciences	2.702 (0.177)		2.991 (0.163)		
Business	3.197 (0.139)		3.428 (0.131)		
Undecided	2.815 (0.105)		2.949 (0.093)		
ACT Math Score Interacted with Major Group:					
Biological, Mathematical, Physical, & Health Sciences	0.035 (0.007)	0.084 (0.006)		0.091 (0.006)	
Liberal Arts	0.022 (0.008)	0.074 (0.007)		0.076 (0.006)	
Engineering and Computer Science	0.041 (0.008)	0.090 (0.008)		0.092 (0.007)	
Education	0.017 (0.009)	0.073 (0.008)		0.082 (0.007)	
Social Sciences	0.034 (0.010)	0.080 (0.009)		0.090 (0.008)	
Business	0.040 (0.007)	0.084 (0.007)		0.090 (0.006)	
Undecided	0.035 (0.006)	0.085 (0.005)		0.088 (0.005)	
ACT Reading Score Interacted with Major Group:					
Biological, Mathematical, Physical, & Health Sciences	-0.003 (0.005)	0.013 (0.005)			0.043 (0.004)
Liberal Arts	-0.012 (0.006)	0.003 (0.006)			0.031 (0.005)
Engineering and Computer Science	-0.010 (0.006)	0.003 (0.006)			0.035 (0.005)
Education	-0.006 (0.007)	0.016 (0.006)			0.042 (0.006)
Social Sciences	0.004 (0.007)	0.017 (0.007)			0.046 (0.006)
Business	-0.004 (0.006)	0.011 (0.006)			0.041 (0.005)
Undecided	-0.018 (0.004)	0.005 (0.004)			0.036 (0.004)

Notes: All models include cohort and university-by-major cell fixed effects. Standard errors included in parentheses. Major-group details:

Biological, Mathematical, Physical, & Health Sciences includes: Agricultural sciences; Natural resources; Biological sciences; Mathematics and statistics; Physical sciences; and Health professions

Liberal Arts includes: Architecture; Ethnic and gender studies; Communications and journalism; Foreign languages; English; Liberal arts, general studies, and humanities; Parks and leisure studies; Philosophy and religious studies; Visual and performing arts; and History

Engineering and Computer Science includes: Engineering; Engineering technologies; and Science technologies

Education includes: Education

Social Sciences includes: Family and consumer sciences; Legal studies; Psychology; Homeland security and law and enforcement; Public administration; and Social sciences

Business includes: Business, management, marketing

*** $p<0.01$, ** $p<0.05$

Table 3. Variance of Student-Level Alignment to University-by-Major Cells with Observed and Counterfactual Sorting Conditions.

	Variance of $M_{i,jm}$
Observed	0.50
Counterfactual Scenarios	
Global Random Assignment	0.93
Global *AI*-Sorting	0.22
Random Assignment Conditional on Initial University	0.65
AI-Sorting Conditional on Initial University	0.30

Notes: This table reports on the system-wide variance of observed and counterfactual academic alignment, measured by the difference between students' own academic preparation and the PPI of the entering cell. See text for description of counterfactual scenarios.

Table 4. The Explanatory Power of High Schools over the PPI of Student Placements.

	Cell PPI (1)	Cell PPI, Net of University PPI (2)
Coefficient on *AI* variable	0.44 (0.04)***	0.32 (0.06)***
Total Model R^2	0.406	0.113
Partial R^2 Attributable to High School Fixed Effects	0.104	0.017

Note: Standard errors clustered by university-by-major cell and high school are included in parentheses. Cell PPI and the individual academic index are normalized such that estimates can be interpreted as mapping a one-standard-deviation move in a covariate to one standard deviation of the true distribution of PPI.
*** $p<0.01$

Table 5. Results from High School Covariate Models, Cell PPI.

	(1)	(2)	(3)	(4)	(5)
Academic Index	0.43	0.43	0.44	0.44	0.44
	(0.04)***	(0.04)***	(0.04)***	(0.04)***	(0.04)***
% HS Minority	0.02			0.00	0.01
	(0.02)			(0.02)	(0.02)
% HS FRL		-0.08		-0.03	-0.03
		(0.01)***		(0.01)***	(0.01)***
Zip % Less than BA			-0.14	-0.13	-0.12
			(0.01)***	(0.01)***	(0.01)***
Basic HS Controls					X
R-squared	0.34	0.35	0.37	0.37	0.38

Notes: Standard errors clustered by university-by-major cell and high school are included in parentheses. The basic high school characteristics included in column (5) are indicators for urbanicity (urban, suburban, town, rural, missing), schools size, and the three geographic context variables: distance to the nearest university in miles, university-level PPI of the nearest university, and the interaction between the two. Cell PPI, the academic index, Pct Minority, Pct Free/Reduced Price Lunch, and Zip Pct Less than BA are all normalized such that estimates can be interpreted as mapping a one-standard-deviation move in the covariate to one standard deviation of the true distribution of PPI.
*** p<0.01, ** p<0.05, * p<0.10

Table 6. Results from High School Covariate Models, Cell PPI Net of University PPI.

	(1)	(2)	(3)	(4)	(5)
Academic Index	0.31	0.31	0.31	0.31	0.32
	(0.06)***	(0.06)***	(0.06)***	(0.06)***	(0.06)***
% HS Minority	-0.01			-0.01	-0.02
	(0.02)			(0.02)	(0.02)
% HS FRL		0.01		0.01	0.00
		(0.01)		(0.01)	(0.01)
Zip % Less than BA			0.01	0.01	0.00
			(0.02)	(0.02)	(0.02)
Basic HS Controls					X
R-squared	0.10	0.10	0.10	0.10	0.10

Notes: Standard errors clustered by university-by-major cell and high school are included in parentheses. The basic high school characteristics included in column (5) are indicators for urbanicity (urban, suburban, town, rural, missing), schools size, and the three geographic context variables: distance to the nearest university in miles, university-level PPI of the nearest university, and the interaction between the two. Cell PPI, the academic index, Pct Minority, Pct Free/Reduced Price Lunch, and Zip Pct Less than BA are all normalized such that estimates can be interpreted as mapping a one-standard-deviation move in the covariate to one standard deviation of the true distribution of PPI.
*** p<0.01, ** p<0.05, * p<0.10

Table 7. The Explanatory Power of High Schools over the PPI of Student Placements: Split Sample Based on Above- and Below-Median AI.

	Above-Median-AI Students		Below-Median-AI Students	
	Cell PPI	Cell PPI net of Univ PPI	Cell PPI	Cell PPI net of Univ PPI
	(1)	(2)	(3)	(4)
Coefficient on *AI* variable	0.78	0.79	0.23	0.05
	(0.07)***	(0.18)***	(0.03)***	(0.03)
Total Model R^2	0.364	0.151	0.190	0.024
Partial R^2 Attributable to High School Fixed Effects	0.114	0.026	0.127	0.023

Note: Standard errors clustered by university-by-major cell and high school are included in parentheses. Cell PPI and the individual academic index are normalized such that estimates can be interpreted as mapping a one-standard-deviation move in a covariate to one standard deviation of the true distribution of PPI.
*** p<0.01, ** p<0.05, * p<0.10

Table 8. Results from High School Covariate Models: Split Sample Based on Above- and Below-Median AI.

	Above-Median-AI Students		Below-Median-AI Students	
	Cell PPI	Cell PPI net of Univ PPI	Cell PPI	Cell PPI net of Univ PPI
	(1)	(2)	(3)	(4)
Academic Index	0.79	0.79	0.23	0.04
	(0.07)***	(0.18)***	(0.03)***	(0.03)
% HS Minority	0.03	-0.02	-0.01	-0.02
	(0.01)*	(0.02)	(0.02)	(0.02)
% HS FRL	-0.04	0.01	-0.02	0.00
	(0.01)***	(0.01)	(0.01)**	(0.01)
Zip % Less than BA	-0.12	-0.02	-0.12	0.03
	(0.01)***	(0.02)	(0.02)***	(0.02)
Basic HS Controls	X	X	X	X
R-squared	0.33	0.13	0.15	0.01

Notes: Standard errors clustered by university-by-major cell and high school are included in parentheses. The basic high school characteristics included in column (5) are indicators for urbanicity (urban, suburban, town, rural, missing), schools size, and the three geographic context variables: distance to the nearest university in miles, university-level PPI of the nearest university, and the interaction between the two. Cell PPI, the academic index, Pct Minority, Pct Free/Reduced Price Lunch, and Zip Pct Less than BA are all normalized such that estimates can be interpreted as mapping a one-standard-deviation move in the covariate to one standard deviation of the true distribution of PPI.
*** p<0.01, ** p<0.05, * p<0.10

Table 9. Alternative Academic Index without High School Class Rank

	Cell PPI		Cell PPI net of Univ PPI	
	(1)	(2)	(3)	(4)
Academic Index	0.52	0.52	0.36	0.36
	(0.05)***	(0.05)***	(0.06)***	(0.06)***
% HS Minority	0.06	0.06	0.04	0.02
	(0.02)***	(0.02)***	(0.02)**	(0.02)
% HS FRL	-0.02	-0.01	0.01	0.02
	(0.01)*	(0.01)	(0.01)	(0.01)
Zip % Less than BA	-0.04	-0.02	0.07	0.08
	(0.01)***	(0.01)**	(0.02)***	(0.02)***
Basic HS Controls		X		X
R-squared	0.37	0.37	0.12	0.13

Notes: Standard errors clustered by university-by-major cell and high school are included in parentheses. The basic high school characteristics included in column (5) are indicators for urbanicity (urban, suburban, town, rural, missing), schools size, and the three geographic context variables: distance to the nearest university in miles, university-level PPI of the nearest university, and the interaction between the two. Cell PPI, the academic index, Pct Minority, Pct Free/Reduced Price Lunch, and Zip Pct Less than BA are all normalized such that estimates can be interpreted as mapping a one-standard-deviation move in the covariate to one standard deviation of the true distribution of PPI.

*** p<0.01, ** p<0.05, * p<0.10

Appendix A
Supplementary Tables

Appendix Table A.1: Summary Statistics for Student and High School Characteristics in the Sample

	Mean	SD
Students in the sample		
High School Percentile Class Rank	0.72	0.21
ACT Math Score	22.63	4.76
ACT Reading Score	24.38	5.51
White Male	0.39	0.49
African American Male	0.02	0.15
Asian Male	0.01	0.09
Hispanic Male	0.01	0.07
Other Race Male	0.01	0.11
White Female	0.49	0.50
African American Female	0.04	0.19
Asian Female	0.01	0.09
Hispanic Female	0.01	0.08
Other Race Female	0.01	0.12
High schools in the sample		
City	0.18	0.38
Suburb	0.38	0.48
Town	0.21	0.41
Rural	0.17	0.38
Locale Missing	0.06	0.24
Number of Students (000)	1.12	0.66
Pct Minority (%)	12.11	16.81
Pct Free or Reduced Price Lunch (%)	10.48	15.35
Zip Pct Less than BA (%)	77.09	13.50
Number of Students	58377	
Number of High Schools	455	
Number of University-by-Major Cells	476	

Notes: Student data are from DHE state administrative records. High school data are taken from the Common Core of Data (CCD). Area information (the share of individuals age-25 and older with at least a bachelor’s degree in the high school’s zip code) comes from the year-2000 United States Census. The high school and local-area averages and standard deviations reported in the table are student weighted.

Appendix Table A.2: Ten Highest and Lowest PPI University-by-Major Cells.

University Level (Selective or Less Selective)	Major	Average AI of Finishers
A. Highest Average AI of Finishers		
Selective University	Nuclear Engineering	4.14
Selective University	Biochemistry	4.12
Selective University	Applied Mathematics	4.12
Selective University	Metallurgical Engineering	4.09
Selective University	Computer Engineering	4.08
Selective University	Industrial Engineering	4.03
Selective University	Agricultural Engineering	4.03
Selective University	Chemical Engineering	4.02
Selective University	Mathematics	4.01
Selective University	Geological Engineering	4.00
B. Lowest Average AI of Finishers		
Less selective University	Journalism	1.78
Less selective University	Business Administration	1.87
Less selective University	Social Sciences, General	1.99
Less selective University	Education, General	2.07
Less selective University	Criminal Justice and Corrections	2.12
Less selective University	Parks, Recreation and Leisure Facilities Management	2.17
Less selective University	English Language and Literature	2.23
Less selective University	Communication and Media Studies	2.23
Less selective University	Fine and Studio Arts	2.24
Less selective University	Criminal Justice and Corrections	2.25

Note: Cells displayed in these tables are restricted to those with at least 10 graduates. University names are masked to preserve anonymity; in total, the cells listed in the table are spread across six of the thirteen universities in the system. “Selective” universities are those with an undergraduate profile considered “more selective” or “selective” in the 2015 Carnegie Classifications of Higher Education. “Less selective” universities in this table are universities with undergraduate profiles that are not considered as selective as “selective” colleges. See http://carnegieclassifications.iu.edu.

Appendix Table A.3. Index Parameters from Primary and Alternative Specifications for the Index, Sparse Version that Does Not Allow the Predictive Power of Pre-Entry Qualifications to Vary by *G*.

	Index Parameters		Additional Specifications		
	(1)	(2)	(3)	(4)	(5)
HS Class Percentile Rank	3.12 (0.06)***		3.31 (0.05)***		
ACT Math Score	0.03 (0.00)***	0.08 (0.00)***		0.09 (0.00)***	
ACT Reading Score	-0.01 (0.00)***	0.01 (0.00)***			0.04 (0.00)***

Note: All models include cohort and university-by-major cell fixed effects. Standard errors included in parentheses.
*** p<0.01, ** p<0.05

Appendix Table A.4. The Explanatory Power of High Schools over the PPI of Student Placements. Analysis Based on Sparse Versions of Equations (1) and (2) that Do Not Allow the Predictive Power of Pre-Entry Qualifications to Vary by *G*.

	Cell PPI (1)	Cell PPI, Net of University PPI (2)
Coefficient on *AI* variable	0.33 (0.03)***	0.13 (0.04)***
Total Model R^2	0.269	0.029
Partial R^2 Attributable to High School Fixed Effects	0.113	0.014

Note: Standard errors clustered by university-by-major cell and high school are included in parentheses. Cell PPI and the individual academic index are normalized such that estimates can be interpreted as mapping a one-standard-deviation move in a covariate to one standard deviation of the true distribution of PPI.
*** p<0.01

Appendix Table A.5. Results from High School Covariate Models. Analysis Based on Sparse Versions of Equations (1) and (2) that Do Not Allow the Predictive Power of Pre-Entry Qualifications to Vary by *G*.

	Cell PPI	Cell PPI net of Univ PPI
Academic Index	0.34	0.12
	(0.03)***	(0.04)***
% HS Minority	-0.01	-0.02
	(0.02)	(0.02)
% HS FRL	-0.03	0.01
	(0.01)**	(0.01)
Zip % Less than BA	-0.14	0.03
	(0.01)***	(0.02)
Basic HS Controls	X	X
R-squared	0.23	0.02

Notes: Standard errors clustered by university-by-major cell and high school are included in parentheses. The basic high school characteristics included in column (5) are indicators for urbanicity (urban, suburban, town, rural, missing), schools size, and the three geographic context variables: distance to the nearest university in miles, university-level PPI of the nearest university, and the interaction between the two. Cell PPI, the academic index, Pct Minority, Pct Free/Reduced Price Lunch, and Zip Pct Less than BA are all normalized such that estimates can be interpreted as mapping a one-standard-deviation move in the covariate to one standard deviation of the true distribution of PPI.

*** p<0.01, ** p<0.05, * p<0.10

Appendix Table A.6: Sensitivity of Primary Findings (Tables 5 & 6, Column 5) to Replacing the Linear AI Control with a 20-Bin AI Control Set.

	Cell PPI	Cell PPI net of Univ PPI
	(1)	(2)
% Minority	0.01	-0.02
	(0.02)	(0.01)
% FRL	-0.03	0.00
	(0.01)***	(0.01)
Zip % Less than BA	-0.12	0.01
	(0.01)***	(0.02)
Basic HS Controls	X	X
R-squared	0.43	0.15

Notes: Standard errors clustered by university-by-major cell and high school are included in parentheses. The basic high school characteristics included in column (5) are indicators for urbanicity (urban, suburban, town, rural, missing), schools size, and the three geographic context variables: distance to the nearest university in miles, university-level PPI of the nearest university, and the interaction between the two. Cell PPI, Pct Minority, Pct Free/Reduced Price Lunch, and Zip Pct Less than BA are all normalized such that estimates can be interpreted as mapping a one-standard-deviation move in the covariate to one standard deviation of the true distribution of PPI. Students are divided into twenty equal-sized bins based on their AI values and we control for the AI bins (coefficients not displayed) in place of the linear AI control used in the main text. This allows for a flexible, highly non-linear relationship between AI and the university-by-major placement PPI, but has no bearing on our findings qualitatively.

*** p<0.01, ** p<0.05, * p<0.10

Table A.7. The Explanatory Power of High Schools over the PPI of Student Placements: Alternate Outcomes Used in Equation (1).

	Outcome: Graduation in 4 years		Outcome: Graduation in 6 years		Outcome: First-year GPA	
	Cell PPI	Cell PPI net of Univ PPI	Cell PPI	Cell PPI net of Univ PPI	Cell PPI	Cell PPI net of Univ PPI
	(1)	(2)	(3)	(4)	(5)	(6)
Coeff. on *AI* Variable	0.49	0.39	0.44	0.32	0.48	0.35
	(0.05)***	(0.07)***	(0.04)***	(0.06)***	(0.04)***	(0.07)***
Total Model R^2	0.433	0.153	0.414	0.113	0.461	0.133
Partial R^2 Attributable to High School Fixed Effects	0.092	0.019	0.106	0.017	0.095	0.018

Note: These results compare to the findings in Table 4. Standard errors clustered by university-by-major cell and high school are included in parentheses. Cell PPI and the individual academic index are normalized such that estimates can be interpreted as mapping a one-standard-deviation move in a covariate to one standard deviation of the true distribution of PPI.

Table A.8. Results from High School Covariate Models: Alternate Outcomes Used in Equation (1).

	Outcome: Graduation in 4 years		Outcome: Graduation in 6 years		Outcome: First-year GPA	
	Cell PPI	Cell PPI net of Univ PPI	Cell PPI	Cell PPI net of Univ PPI	Cell PPI	Cell PPI net of Univ PPI
	(1)	(2)	(3)	(4)	(5)	(6)
Academic Index	0.50	0.38	0.44	0.32	0.49	0.35
	(0.05)***	(0.07)***	(0.04)***	(0.06)***	(0.04)***	(0.07)***
% HS Minority	0.01	-0.00	0.01	-0.02	0.02	-0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)
% HS FRL	-0.03	-0.01	-0.03	0.00	-0.03	0.01
	(0.01)***	(0.01)	(0.01)***	(0.01)	(0.01)***	(0.01)
Zip % Less than BA	-0.12	-0.00	-0.12	0.00	-0.11	0.02
	(0.01)***	(0.02)	(0.01)***	(0.02)	(0.01)***	(0.02)
Basic HS Controls	X	X	X	X	X	X
R-squared	0.41	0.14	0.39	0.10	0.44	0.12

Notes: These results compare to results from the full specification in Tables 5 and 6. Standard errors clustered by university-by-major cell and high school are included in parentheses. The basic high school characteristics included in column (5) are indicators for urbanicity (urban, suburban, town, rural, missing), schools size, and the three geographic context variables: distance to the nearest university in miles, university-level PPI of the nearest university, and the interaction between the two. Cell PPI, the academic index, Pct Minority, Pct Free/Reduced Price Lunch, and Zip Pct Less than BA are all normalized such that estimates can be interpreted as mapping a one-standard-deviation move in the covariate to one standard deviation of the true distribution of PPI.
*** p<0.01, ** p<0.05, * p<0.10

Appendix Table A.9: Sensitivity Analysis: Dropping Cells without Finishers (N = 39549).

	Cell PPI	Cell PPI net of Univ PPI
	(1)	(2)
Academic Index	0.45	0.45
	(0.04)***	(0.06)***
% Minority	0.01	-0.03
	(0.02)	(0.02)
% FRL	-0.04	-0.00
	(0.01)***	(0.01)
Zip % Less than BA	-0.13	-0.01
	(0.01)***	(0.02)
Basic HS Controls	X	X
R-squared	0.33	0.13

Notes: Standard errors clustered by university-by-major cell and high school are included in parentheses. The basic high school characteristics included in column (5) are indicators for urbanicity (urban, suburban, town, rural, missing), schools size, and the three geographic context variables: distance to the nearest university in miles, university-level PPI of the nearest university, and the interaction between the two. Cell PPI, the academic index, Pct Minority, Pct Free/Reduced Price Lunch, and Zip Pct Less than BA are all normalized such that estimates can be interpreted as mapping a one-standard-deviation move in the covariate to one standard deviation of the true distribution of PPI.

*** p<0.01, ** p<0.05, * p<0.10