6-1-2015

How Can We Accurately Measure Whether Students are Gaining Relevant Outcomes in Higher Education?

Tatiana Melguizo
University of Southern California

Gema Zamarro
University of Arkansas, Fayetteville, gzamarro@uark.edu

Tatiana Velasco
Universidad de los Andes, Colombia

Fabio Sanchez
Universidad de los Andes, Colombia

Citation

Melguizo, T., Zamarro, G., Velasco, T., & Sanchez, F. (2015). How Can We Accurately Measure Whether Students are Gaining Relevant Outcomes in Higher Education?. *Education Reform Faculty and Graduate Students Publications.* Retrieved from https://scholarworks.uark.edu/edrepub/46

How can we Accurately Measure whether Students are Gaining Relevant Outcomes in Higher Education?

Tatiana Melguizo*
University of Southern California

Gema Zamarro**
University of Arkansas

Tatiana Velasco and Fabio J. Sanchez***
University of Los Andes

EDRE Working Paper 2015-03
Last Updated June 2015
Please Do Not Cite Without Permission

Abstract

The main objective of this study is to empirically test a number of theory-based models (i.e. fixed effects (FE), random effects (RE), and aggregated residuals (AR)) to measure both, the generic knowledge as well as the degree attainment rates and early labor outcomes, gained by students in different programs and institutions in higher education. There are four main findings: First, the results of the paper confirm the need of using models that address the issue of student selection into programs and institutions in order to avoid biased estimates. Second, our findings provide suggestive evidence in favor of using FE models. Third, the results also illustrate the need to use appropriate statistical corrections (e.g., Heckman type selection models) to also address the issue related to students dropping out of college. Finally, our findings confirm our hypotheses that rankings of specific college-program combinations change depending on different educational and labor outcome measures considered. This finding emphasizes the need to use complementary indicators related to the mission of the specific post-secondary institutions that are being ranked. The results of this paper illustrate the importance of validating empirical models intended to rank college-program contributions according to a number of educational and early labor market outcomes. Finally, given the sensitivity of the models to different model specifications, it is not clear that they should be used to make any high-stakes decisions in higher education. They could, however, serve as part of a broader set of indicators to support programs and colleges as part of a formative evaluation.

Acknowledgements

This study was supported by a small grant by the Spencer Foundation (2015-00027). Opinions are those of the authors alone and do not necessarily reflect those of the granting agency or of the authors' home institutions. We would like to thank Matthew Johnson and conference participants at the 2015 AEFP 40th Annual Conference as well as the participants of the 2015 XXIV Economics of Education Association for valuable feedback on a previous version of this paper.

*Rossier School of Education, 3470 Trousdale Parkway, WPH 702 G, Los Angeles, CA, 90089. Phone: (213) 740 3635 Fax: (213) 740 3889 Email: melguizo@usc.edu
**Department of Education Reform, 219-B Graduate Education Building, Fayetteville, AR 72701. Phone: (479) 575 7024 Email:gzamarro@uark.edu
*** School of Economics, Universidad de los Andes. Calle 19A No 1-37 Este. Bloque W Phone: (571) 339 49 49 / Fax: (571) 332 44 92. email: fasanche@uniandes.edu.co; tatianavelasco.ro@gmail.com

Recent estimates suggest that the U.S. has lost its leadership in terms of postsecondary degree attainment in the world and that its adult population currently has average educational levels compared to counterparts in other developed countries (OECD, 2013). The Organization for Economic Cooperation and Development (OECD) recently released the findings of the Program for the International Assessment of Adult Competencies (PIAAC) that tested the numeracy, literacy, and problem-solving skills in technology rich environments of adults (16-65 years old) in 23 countries. The adults in the U.S. ranked near the middle in literacy and near the bottom in skills with numbers and technology (Pérez-Peña, 2013). These new results are worrisome as they suggest that not only are the overall degree attainment rates in the U.S. lower, but also that adults are not gaining the knowledge and skills necessary to succeed in the work place.

The Obama administration has attempted to address this issue through the American Graduation Initiative, and more recently through a proposal to tie financial aid to a college ranking system (Field, 2013; Stratford, 2015). The proposal includes creating a college rating system based on measures of access, affordability, and student outcomes, and to allocate student aid based on these ratings. This proposal has received mixed reactions. The main criticism is that the U.S. lacks a data system or set of credible indicators that can be used to accurately measure and evaluate a number of relevant educational outcomes (i.e., student learning outcomes (SLOs), persistence, degree attainment, dropout rates, and early earnings) provided by different types of post-secondary institutions. The reality is that the pressure for making colleges and universities more accountable is only going to increase. Academics and college officials are responding to this mounting pressure, by developing a complex and thoughtful set of theory-based models that can be tested empirically and used as part of a formative evaluation of higher education institutions. The OECD, through its Assessment of Higher Education Learning Outcomes (AHELO)[1] project, designed a multi-dimensional, inter-disciplinary, and cross-cultural system to compare the SLOs of students attending colleges in higher education institutions of member countries. The goal included creating a set of reliable and validated instruments to measure the student gains after studying in postsecondary institutions in OECD countries. Unfortunately, the methodological challenges implicit in creating these types of indicators, as well as the lack of financial support and political will, resulted in the postponement of this important project.

A number of countries in other parts of the world have developed centralized assessment and accountability systems along with instruments to measure SLOs in higher education. Student learning outcomes have been traditionally measured using tests developed to measure the generic knowledge (i.e., critical thinking, problem solving, and civic education) or subject-specific knowledge (i.e., Engineering or Sociology) in a specific program of study. Specifically, countries like Australia, Brazil, and Colombia have already created centralized accountability systems to evaluate and rank higher education institutions that include different types of indicators (Coates, 2009; Melguizo & Wainer, 2014). This paper capitalizes on the rich data from instruments built in Colombia in the past two decades to empirically test a number of theory-based models to measure the generic knowledge[2] gained by students in different programs in higher education. The following research questions guide this study: How can we measure whether postsecondary institutions are providing generic knowledge to students enrolled in different programs in higher

[1] The main goal of AHELO is to create a multi-dimensional, inter-disciplinary, cross-cultural, and comprehensive system to evaluate whether students were indeed learning valuable knowledge and skills. For a more detailed description of the project and preliminary results of the pilot program see OECD (2014).

[2] Colombia uses a college-exit exam, SABER PRO (described in more detail below) that measures both generic and subject-specific knowledge. Whereas students in all programs have taken the test to measure the generic component, this has not been the case for the subject-specific test. Unlike the generic test that was adapted from the Collegiate Learning Assessment (Klein, Benjamin, Shavelson, & Bolus, 2007; Rossefsky-Saavedra & Saavedra, 2011), the subject specific tests have been developed sequentially for the different programs. For the purpose of this project, we decided to focus on the generic component of the SABER PRO.

education? How do college quality[3] measures based on generic knowledge compare with measures based on college contribution to graduation probability and success in the labor market? This study contributes to the literature by proposing and empirically testing models to measure student gains both generic knowledge and other relevant student educational outcomes such as graduation and early labor outcomes. These two pieces of information will allow us to study whether the most selective institutions might be adding relatively low value in terms of general knowledge and skills, but they might be adding value because these students have higher graduation rates and higher success in the labor market. This information is crucial because if evaluation systems focus only on traditional outcomes instead of both a generic knowledge proxy for student learning outcomes and relevant short- and long-term outcomes, institutions might have incentives to become more selective, limit their curriculum, or reduce the focus on programs with lower graduation rates like Science, Technology, Engineering, and Mathematic (STEM).

Conceptual Framework and Literature Review

The idea of creating a centralized assessment and evaluation system of higher education institutions in the U.S. is a relatively recent phenomenon (Field, 2013; Spellings, 2006) and consequently the scholarship related to conceptualizing models for assessment and evaluation is relatively new. For this reason, we use models proposed in other countries that have more centralized governance structures and have successfully engaged in designing complex nationwide centralized assessment and evaluation systems in higher education (Coates, 2009; Melguizo & Wainer, 2014).

Coates (2009) proposed a comprehensive model that includes measuring the "value-

[3] Quality is a very elusive concept in higher education (Melguizo, 2008). In this paper we do not attempt to measure the quality but rather any gains in general knowledge and skills as measured by valid and reliable tests.

added" (VA) by higher education institutions in Australia. He argues that measuring learning at the higher education level is a complex but vital issue. Coates lists and describes four approaches that can be combined into a single model and used to measure the "quality" and "value-added" of higher education institutions: 1) computation of value-added estimates by comparing predicted against actual performance using data from entrance tests and routine course assessments, 2) comparison of outcomes between objective assessments administered to cohorts in their first and later years of study, 3) comparison of first and later years student engagement, and 4) feedback on graduate skills provided by employers. In this paper we focus on the second indicator and propose and empirically test theory-based models that attempt to provide measures of the generic knowledge gained by students in higher education, as well as other relevant outcomes.

We proceed to describe the scholarship that has been developed in recent years and that relates to attempts to measure SLOs in higher education. Cunha and Miller (2014) proposed the Input-Adjusted Outcomes Measure (IAOMs) model as a tool to provide policy makers information on the educational outcomes produced by a number of post-secondary institutions in their states. The idea was to create a comprehensive dataset using information collected by K-12, higher education, and employment offices, and run a regression controlling for as many observable characteristics as possible. The regressions would provide estimates for specific outcomes such as persistence, graduation, and earnings, which can be used to rank and compare institutions. This model is appealing because of its simplicity. The main problem, acknowledged by the authors, is that the estimates might be biased because of students' self-selection into specific colleges (e.g. students deliberately choose institutions where they think would have better academic outcomes).

Most recently a handful of studies attempted to measure the knowledge and skills gained by students in college (See, Arum & Roksa, 2011; Barrera-Osorio & Bayona-Rodriguez, 2014; Domingue, Morales, Shavelson, Wiley, Molina, & Mariño; 2014; Melguizo & Wainer, 2014; Rossefky-Saavedra & Saavedra, 2011; Saavedra, 2009; Steedle, 2012).

In their book “Academically Adrift,” Richard Arum and Josipa Roksa (2011) attempted to measure whether students in the U.S. were learning valuable skills in higher education. They used the Collegiate Learning Assessment (CLA) instrument to test over 2,000 freshmen in 24 institutions. The authors concluded that 45 percent of students “did not demonstrate any statistically significant improvement in learning” during the first two years of college (p. 121). The main limitation of this study is the lack of acknowledgment or use of any statistical corrections to address the problem of selection of students into certain institutions or programs. In other words given that students self-select into colleges and into programs/majors, the educational outcomes resulting from comparing students at different types of institutions are the result of the pre-entry characteristics of these students (i.e., academic preparation and motivation), and not what institutions are indeed providing to these students. All the studies described below have acknowledged this problem and used a number of empirical models to avoid getting biased estimates.

Melguizo and Wainer (2014) conducted a descriptive study to provide some initial measures related to gains in SLOs in higher education. They used the ENADE, an instrument designed to measure learning growth in Brazil that measures gains in terms of general and subject-specific areas. The fact that Brazil administered a college-level exam, the ENADE, to both freshmen and senior students the same year provided a unique opportunity to get a first approximation of the generic and subject-specific area knowledge gained in different programs. The results suggested that, on average, students in the three different categories of programs (e.g., Science, Technology, Engineering, and Mathematics (STEM), Social Sciences, and Biological Sciences) were gaining valuable general and subject area knowledge. The results showed that the gains in the subject area were of a larger magnitude than those in the general knowledge component of the test. This study noted the problem of selection of students into colleges, and used propensity score matching methods to address it.

Rosefky-Saavedra & Saavedra (2011) used information from Colombia for a cohort of students in 2009 to estimate value-added models in higher education. They concluded that relative to observationally similar high school graduates, students in the last year of college scored about half of a standard deviation higher, with statistically significant higher scores on every component of the generic test. This is one of the few studies to attempt to correct for the selection of students into colleges. Nonetheless, as noted by the authors, this study suffers from a number of limitations. First, the study used data from a pilot study, which is problematic, given that the students who chose to take the test are not representative of the average student population, limiting the external validity of the findings. Second, at the time of the study, the SABER PRO was not compulsory, with no real consequences attached to performance. Third, the data from the pilot study relied on a different cohort of students. Ideally, any model that attempts to measure the knowledge and skills should administer the same exam twice to the same cohort of students: once before starting their first year, and once again as seniors. Finally, there is also the problem of possible maturation bias, so students who were not enrolled in college would also be expected to experience gains in terms of generic knowledge and skills as they age. Despite these limitations, this study is one of the first to provide robust initial estimates of SLOs in a higher education context.

Saavedra (2009) used Colombian data to measure the return of attending more selective colleges and universities in terms of learning and earnings. He used a sample of college graduates between 2001 and 2007 and a regression discontinuity design (RDD), to estimate a number of effects for the students at the "cutoff." He concluded that the students who were admitted to the most selective university in Colombia had substantial gains in learning, about 0.2 of a standard deviation, a higher probability of being employed, and higher earnings compared to the students right below the cutoff who attended less selective institutions. He also found that low SES students also had the largest gains on college-exit exam scores. Most of the limitations described above related to the Rossefsky-Saavedra & Saavedra study (2011) also apply to this paper.

Barrera-Osorio and Bayona-Rodriguez (2014) used data for over 80,000 students that applied to a Colombian private elite institution to explore the impact of the "*quality"* of the postsecondary institution attended on a number of educational outcomes such as: probability of enrollment, course failure rate, dropout rate, results on a college-exit exam, probability of finding a job in the first year after graduation, and salary. The authors wanted to test whether attending an elite institution translated in gains in the outcomes described above as predicted by human capital theory (Becker, 1962), or as predicted by the signaling theory (Arrow, 1973), where one's educational history signals to the market one's personal characteristics. The authors used a fuzzy regression discontinuity design (RDD) and compared the outcomes of individuals right above and right below the admission cutoffs. The assumption is that these individuals should be similar in terms of observed and unobserved characteristics, and any differences in the outcomes should be related to the "quality" of the institution attended. The results suggested that compared to those who attended different or less elite institutions, individuals who attended the elite institution had higher outcomes in terms of relatively higher enrollment and graduation rates (although results varied by program), and higher earnings. Students who enrolled in the elite institution had a slightly higher probability of failing a course, no clear difference in terms of dropping out of college (with the exception of students in Engineering, who had a lower rate of drop out), and no difference in knowledge and skill gains measured in terms of the difference between gains in the college-exit exam, SABER PRO, and the high school entrance exam SABER 11.[4]

The authors conclude that their results seem to support the predictions of signaling theory: despite a lack of observed gains in knowledge and skills, the graduates of the elite university are reaping higher earnings. This paper has a number of limitations. First, like Saavedra (2009), this paper uses RDD to compare individuals in the "treatment" group (i.e., selective institutions) with individuals in different "control" groups (i.e., selective and non-selective institutions). This is problematic as the control group is composed of individuals who could have attended multiple types of institutions, so we do not really know what the differences in the estimates relate to. Second, the authors included cohorts of students before and after the exam was compulsory, which might have biased the estimates. Finally, unlike Saavedra (2009), the authors decided to use the combined (i.e., generic and subject specific) SABER PRO. These two components are intended to measure very different sets of knowledge and skills and were not designed to be combined. This decision might also help explain the contradictory findings of these two relatively similar papers.

Domingue et al. (2014) use the Colombian college-exit exam SABER PRO to identify the challenges of using VA models in higher education, in particular, as compared to the use of these

[4] It is important to note the discrepancies in the findings related to gains in knowledge as skills between this paper and Saavedra (2009). As mentioned above the positive results of Saavedra's paper might be related to the use of the self-selected sample of individuals who participated in the pilot program of SABER PRO. This also raises the issue of the importance of using not only the appropriate methods but representative data from the student population.

models in the K-12 setting. The authors clearly state that the results of VA models cannot be interpreted as causal,[5] and they recommend making a number of operational decisions when using VA models in higher education, such as: 1) defining the treatment and control; the authors invite researchers to consider multiple potential control groups for a college, such as comparing estimates with either the "average" college, another college that the individual was predicted to attend, or not attending college; 2) defining the unit of analysis after defining the treatment: the challenge in higher education is that it is more difficult to define a unique treatment-- while students choose specific majors, they take courses from different departments that contribute to their general knowledge and skills; 3) defining outcomes: the challenge here is to choose between measuring generic versus subject-specific skills; 4) inclusion of covariates: the models are sensitive to the inclusion of particular covariates such as previous academic preparation (i.e., high school-exit exam) and peer effects (i.e., average score on high school-exit exam for a cohort); 5) missing values: estimates may be biased when they are associated with the dropout of students from college; and 6) ability sorting, a clear threat to making causal statements using VA models in higher education. The authors used results from SABER PRO in Colombia to empirically test VA models in higher education, and explore how sensitive the results were to the different model specifications used in terms of: 1) ability sorting, 2) covariates included, 3) choice of outcome (i.e., generic versus subject-specific), and 4) relationship between VA estimates and attrition rates. The authors used longitudinal data for a sample of over 60,000 students who took the SABER PRO exam during 2011 and 2012 and estimated three different types of models. Model 1 estimated the gains in the generic component of SABER PRO for each institution by reference (IBR) unit; controlled for previous scores on SABER 11, the national

[5] The authors list a number of assumptions that need to be met in order to make causal statements (e.g., manipulability, homogeneity, strong ignorable treatment). They then explain the challenges of having these assumptions hold in a setting were students can't be assigned randomly to the treatment.

high school exit exam; and used a random effects model. Model 2 attempted to control for peer-effects by adding a measure of the average SES of the individuals that are part of the IBR unit. Finally, Model 3 used an alternate control for peer effects that consisted of the mean SABER 11 for each IBR unit. The authors' discussions of the empirical findings, in light of the different methodological challenges associated with estimating VA models in the higher education setting, are summarized below.

Ability Sorting. The empirical results of Domingue et al. (2014) suggest that there is indeed ability sorting by program. This is not surprising given that there are some majors such as STEM fields which are more competitive than other majors in, for example, humanities. The authors advocate using models to make comparisons across universities for given programs/majors, but caution that this might result in less precise estimates due to the small sample sizes.

Inclusion of covariates. Domingue et al. (2014) compare the VA estimates in engineering in the case of a very selective program (i.e., very high SABER 11 and SABER PRO scores), and an engineering program with average selectivity. The authors show that the VA estimates for these two programs under Model 1 are similar and place them in the "average" effectiveness group, whereas the VA estimates for Models 2 and 3, which control for peer effects, present a very different picture, with the more selective engineering program presenting much lower effectiveness.

Subject-specific outcomes. The authors compare the results of Models 1 and 3 in terms of the generic and subject-specific parts of SABER 11 in Education and Law programs. They find strong correlations in the estimates of these models, and conclude that switching between generic and subject-specific outcomes in this case had very little consequences in the VA estimates. Consistent with Melguizo and Wainer (2014), they found that VA estimates of subject-specific outcomes explain a much larger proportion of the variability in student scores compared to estimates of generic skill outcomes.

Attrition. The estimates reported by Domingue et al. (2014) suggest higher dropout rates are associated with lower VA estimates in Models 1 and 2, but not in Model 3.

Our paper builds on the work of Domingue et al., (2014) by focusing on a different set of models, fixed-effects models, and conducting sensitivity analyses to check for changes in the model specifications, including corrections for student attrition. Our goal, like theirs, is to contribute to the methodological debate and warn college administrators and policy makers of the perils of making high-stakes decisions based on models that have not been fully validated.

Methodology

Value-added models have their origin in K-12 education and are often used to evaluate teacher or school quality (Briggs, 2012; Buddin & Zamarro, 2009; Chetty, Friedman, & Rockoff, 2011; Gorard, 2008; Kane & Staiger, 2008; Sass, Hannaway, Xu, Figlio, & Feng, 2012). The two main challenges related to estimating VA models in higher education in the U.S. are: 1) having access to a reliable and valid instrument that measures the knowledge and skills gained by students in college, and 2) addressing selection bias from students deliberately choosing to attend certain colleges. Without good predictors of college choice, it is difficult to separate the college's contribution to learning from students' unobserved attributes such as motivation and natural ability.

As mentioned above, Colombia has invested substantial resources in the development of a set of valid and reliable instruments to measure the knowledge and skills gained by students in high school (i.e., the SABER 11 national high school exit exam) and college (i.e., the SABER PRO college exit exam). Specifically, the national level datasets available in this country are ideal for conducting empirical estimations of VA models. First, Colombia has results for each student from instruments designed to measure learning at the end of high school (SABER 11) and college (SABER PRO). Second, Colombian students who want to access postsecondary education apply not only to specific colleges but also to a specific program (e.g., Economics). Each college then selects students into a specific program mainly by looking at the student's SABER 11 scores. Given that we have access to information on the SABER 11 exam, which is the key driving factor of college-program enrollment decisions in Colombia, we are better able to correct for selection of students into universities and programs and to assess the importance of such sorting bias. Only after correcting for this selection bias will we be able to assess whether VA models based on college-exit exams are promising methods to obtain meaningful estimates of the SLO gained by students in different programs in different postsecondary institutions.

SABER 11 and SABER PRO

SABER 11 is a compulsory high school-exit exam in Colombia. This test takes place twice every year (fall and spring) corresponding to two different high school graduation cohorts. As part of the test, socio-economical information of the students is gathered and knowledge in areas such as Mathematics, Physics, Chemistry, Biology, Language, Philosophy, Social Science and English is evaluated. A substantial number of private and public universities in the country use the score in the SABER 11 exam to admit students into selective postsecondary institutions and all of them require the applicant students to have successfully presented the test in order to be considered for admission (Barrera-Osorio, F., & Bayona-Rodríguez, 2014).

SABER PRO[6] is the college-exit exam; since 2009, it has been compulsory for graduation for all students who completed 75% of the college program. It is composed of a

[6] For a more detailed description of SABER PRO see Domingue et al. (2014).

generic and a subject-specific component. The generic part is based on the College Learning Assessment (CLA) and, since 2011, includes four modules: writing, English, reading/critical thinking, and problem solving. The Colombian Institute for Higher Education (ICFES, acronym in Spanish) has been designing the subject-specific exams since 2007. Thus, it has implemented different versions of this part of the test since that year. Nevertheless, all modules of the generic component have been compulsory for students since the second semester of 2011. Also, since 2011 every program in the country has a subject-specific component, in addition to the generic component described above. In this paper we use the combined score of the generic component for all the students who have taken the test since the second semester of 2011, when the generic exam was fully developed and the SABER PRO was a compulsory requirement for graduation.

Datasets used to track students in college and after graduation

In order to identify the semester of entrance, graduation and labor market entrance for each student, we use two different datasets: the System for the Prevention of College Dropout (SPADIES, acronym in Spanish) and the information of the Labor Observatory for Education (OLE by its acronym in Spanish). Both sources are administered by the Colombian Ministry of Education. The first one gathers biannual information on all students who enter higher education, tracking them until they either drop out or graduate. In particular, this dataset provides us with information about the program a student attended, the college s/he attended, whether the student has graduated or drop out and her/his entrance cohort. The second dataset collects annual information for all graduates of higher education in Colombia on employment status, economic sector, and current salary.

Data and Sample

We use a comprehensive dataset that links information collected by the ICFES to that collected by the SPADIES and the OLE. The Colombian Education Ministry linked the information for each student in these datasets and created a unique identification that allows us to track each student from the time they take SABER 11 until they appear in the OLE data. Our dataset includes information for multiple cohorts of students, which we use to conduct the estimates for the outcomes of interest. In the case of SABER PRO, we focus on students who took the test between 2011-2 and 2012-2 (2011-2 refers to students who took the test in the second semester of 2011 and 2012-2 refers to students who took the test in the second semester of 2012) and who enrolled in university between 2006-1 and 2008-2.[7] In the case of graduation, we focus on cohorts of students who enrolled between 2003-1 and 2007-1 and we follow them to 2012, which would give the last cohort a minimum of five years to graduate. For the analyses related to early labor market outcomes, the focus is on cohorts of students who graduated between 2005-1 and 2011-2, for which we have data until 2013-1. This data is censored for the late cohorts, since an individual who graduated in 2005 will be in the dataset for a longer period of time than the individual who graduated in 2011. However, given that most students get a job within two years after graduation, we think that this relatively shorter period is appropriate to observe the early labor outcomes of interest in this study.

Our sample for each outcome in the described cohorts is composed as follows: for SABER PRO we observe 74,421 students. They represent 197 colleges with students in one of the following program categories: agriculture and veterinary, arts, education, health, social sciences and humanities, economics and business, engineering and architecture, and math and natural sciences. For the case of graduation, we observe 245,358 students. They represent 195 colleges with students in the program categories described above. Finally, for early labor market

[7] These are the entering cohorts of students who would be expected to take the SABER-PRO exam in the selected period 2011-2 to 2012-2.

outcomes, we observe 146,446 graduated students from 229 different colleges and find 99,790 of them with a formal job status.[8]

Models

We estimate value-added contributions by college-program to students' outcomes using versions of the following equation:

$$OUTCOME_{itcp} = \beta_1 X_i + \beta_2 SABER11_i + \beta_3 \overline{SABER11_{\imath cp}} + \delta_{cp} + \varepsilon_{itcp} \text{ (1)}$$

Where $OUTCOME_{itcp}$ denotes either standardized results in SABER PRO by graduating cohort and year, or measures of graduation or employment of student *i* of cohort *t* that graduates from college *c* and program *p*. X_i contains the following relevant student demographic information that determines both the outcome of interest and selection in specific colleges and programs: student gender, parental socio-economic status and mother's education. With the student's test results in SABER 11, these variables allow us to control for selection bias due to students' choices of college and program. We also control for the average SABER 11 scores for each student's entering cohort in a given college and program. In that way we control for differential peer cohort qualities and obtain value-added college contributions purged of cohort effects. Finally, SABER PRO cohort dummies in regressions for SABER PRO, entrance to college cohort effects for regressions on graduation, and graduating cohort effects for analysis of labor market participation are in the specification to control for any remaining cohort effects. Finally, δ_{cp} , our main parameters of interest, identifies college by program effects measuring

[8] A formal job entails the payments of contributions and taxes to the Colombia's social security system in particular for pensions and health both by the part of the employer and the worker.

how much on average students in a given college and program perform above those in other colleges and programs who have similar characteristics. Note that by estimating models that do not include a constant term, we are able to estimate these effects for all possible college-program combinations in our data, avoiding the problem of having to choose a college-program as reference.

Our preferred specification treats these effects as fixed effects and will then control for any correlation among the college-program effects and our explanatory variables. We believe this to be the most appropriate specification as one would expect that college-program contributions would potentially correlate with the explanatory variables, especially those related to the selection of students into colleges and programs. That is, one would expect that those institutions that contribute more to students' general knowledge, graduation, or labor outcomes might be also those with higher selectivity of students into their college or programs. Not controlling for this potential correlation could lead to biased estimated coefficients, including biased measures of college-program effects.

To assess the importance of controlling for this, we also test for other specifications often used in the context of K-12 education (e.g., random effects and aggregated residual methods). Random effects (e.g., McCaffrey, Lockwood, Koretz, & Hamilton, 2003), and Aggregated Residual Methods (e.g., Kane and Staiger, 2008) are two methods where college-program effects are eliminated from the estimation equation and college-program contributions are obtained as averages of the estimated residuals after controlling for the rest of the covariates in equation (1). By comparing results with these diverse methods, we are able to assess how sensitive our college-program contribution estimates are to alternative specifications and we will be able to compare our estimates with those of Domingue et al. (2014) who used a random effects approach.

In addition, we also estimate models without controlling for the results of SABER 11 or cohort effects ($\overline{SABER11_{tcp}}$). By comparing these estimates with the ones obtained from equation (1) above, we are able to assess the relevance of controlling for selection bias. This paper contributes to the literature beyond the work of Domingue et al (2014), by conducting not only random effect (RE) estimates, but by proposing and empirically testing two additional types of models: aggregated residual (AR) and fixed effects (FE). We agree with Domingue et al (2014) that the assumption implicit in RE models is strong, so we propose a use of FE that relaxes some of the assumptions and enables correlation between the control and explanatory variables. In addition, we not only focus on the VA in terms of knowledge and skills but also include other relevant outcomes such as graduation and early labor outcome measures.

Correction to address the problem of differential attrition due to different dropout rates

An additional complication of computing VA estimates comes from the potential bias due to non-random dropout rates of students and the fact that only students who graduate take the SABER PRO exam. To the extent that the demographic variables and test scores in SABER 11 are important predictors of dropout rates, by controlling for them we would also take into account this potential bias. As an alternative, one could use propensity score weighting methods to guarantee that our resulting graduating cohorts are balanced, within program across colleges, in these observed characteristics.

However, these proposed methods won't control for potential unobservables determining graduating from college that could also be linked to the unobservables determining results in SABER PRO tests. To take this into account, we estimated traditional selection correction models such as in Heckman (1978). To help identify these Heckman's type models, as our exclusion restriction, we added information about what the local unemployment rates in the area of the student's college were at the moment of our last observation of student enrollment. That is, we assume this variable determines dropout decisions but does not affect the results of SABERPRO directly.

Results

We begin by summarizing program-college averages for the outcomes of interest (See Table 1). Looking at the distribution of generic SABER PRO scores, it is clear that the highest scores are attained by students in Math and Natural Sciences programs. We standardized the scores by graduating cohort and year. This means that the average SABER PRO scores in the Math and Sciences programs are almost half a standard deviation higher than the results of students from the same graduating cohort but in other programs who took the SABER PRO exam during the same period.

Engineering and Architecture and Arts also had relatively high SABER PRO scores. The programs with lower scores were Education, Agriculture and Veterinary, Economics and Business, and Social Sciences and the Humanities. It is worth noting the relatively high standard deviations in SABER PRO scores, indicating considerable heterogeneity in student performance across colleges within any given program area. It is also interesting to note that the program-college combinations with higher average SABER PRO scores corresponded with those with the highest SABER 11 scores. Math and Sciences, Engineering and Architecture, and Art had above average SABER 11 scores, while Education had the lowest. These results hint at the selection of students into different program-colleges which could explain part of the observed higher SABER PRO scores described above. Finally, probably the most interesting findings are related to the homogeneity in terms of degree attainment/graduation rates and early labor market outcomes by program. In the case of labor participation, it is interesting to see the high participation rates of student in all programs, with a range of between 57 and 71 percent. Finally, graduation rates ranged from 40 to 58 percent, with the highest program-college combination in Health and Economics and Business.

<<Table 1>>

Aggregated Residuals (AR), Random (RE) and Fixed Effects (FE) Estimates of College-program effects in General Knowledge. Following the specification described in (1), we estimated models using AR, RE, and FE methods and used a Spearman Rank correlation to study the degree by which our estimated program-college rankings obtained through each of these methods were correlated (Tables 2A and 2B). The results in Table 2A show that our estimated rankings of program-college combinations are sensitive to estimate methods for certain model specifications. In particular, when controls for selection (i.e., SABER 11 scores and cohort average SABER 11 scores) are excluded, the three estimation methods return rankings that are relatively similar. We see correlations of above 95 percent for each method combination: AR versus FE, RE versus FE, and AR versus RE.

However, once we start to address the issue of selection by controlling for either SABER 11 scores, average SABER 11 entry cohort peer effects, or both of these controls, the correlations diminish and we see weaker correlations in the case of the AR versus FE methods. We also studied the sensitivity of college-program rankings to model specification within a given method in Table 2B. In this case the most stable rankings are the ones provided by FE models; they are more robust whether controls for selection are considered or not. These estimates provide preliminary empirical evidence that seems to favor the use of FE models as opposed to RE models.

<<Tables 2A and 2B>>

Ranking according to college-program contributions to SABER PRO. We rank the programs using FE models and present detailed descriptive statistics of the distribution of estimated college-program effects before and after controlling for selection, using both the SABER 11 scores and peer effects based on average SABER 11 results for the entry cohort of the student (See Table 3A and Table 3B and Figures 1A and 1B). As clearly illustrated in the figures, once we introduce the controls for selection, the VA gains in the different program-college combinations diminish substantially. Similarly, Table 3A and 3B show that once you control for selection, the VA gains observed in the majority of the programs basically disappear. This is consistent with findings by Domingue et al (2014) and Melguizo & Wainer (2015), who also found very small gains in the generic component of the Brazilian college-exit exam. It is worth noting that for institutions in the top 75th percentile of the distribution, there is evidence of positive contributions. These results suggest that a small number of probably accredited public and private institutions are adding value in terms of increases in generic knowledge and skills.

<<Tables 3A and 3B>>

<<Figures 1A and 1B>>

Ranking according to college-program contributions to Graduation rates. We also rank each program-college related contribution to graduation rates (Tables 4A and 4B), using a Linear Probability Model (LPM) on specification (1) above, where the outcome variable is a dummy variable for observing the student graduating. The results suggest similar average contributions to graduation rates, of around 0.46 to 0.65, for all the different program-college combinations with and without controlling for selection of students.

We are also interested in exploring whether the program-college contributions that were adding the highest value in terms of this outcome were the ones that were also adding value in terms of general knowledge and skills (SABER PRO). We use Spearman Rank correlation coefficients to rank college-programs based on SABER PRO and graduation rates. The goal was to test whether universities that were ranked as adding more value in terms of knowledge and skills were the ones that were also ranked as adding more value in terms of degree attainment. We found that before controlling for selection, there was a considerable correlation in the rankings in these two outcomes in Agriculture, Social Sciences and Humanities, Economics and Business, and Math and Natural Sciences; however, after the control for selection, the correlation becomes almost zero or even reverses the sign. This finding is consistent with Barrera-Osorio and Bayona-Rodríguez (2014) who also found small contributions in terms of increase in generic knowledge, and that institutions were mostly contributing in terms of graduating students and enabling them to get jobs.

<<Tables 4A and 4B>>

Ranking according to college-program contributions to Being Employed in the Formal Sector. We also study college-program contributions to the probability of formal participation in the labor market. As in the case of graduation, we follow a linear probability model on specification (1) above. The results suggest similar high program-college contributions, among graduates, to the probability of being employed for all programs ranging from 0.70 to 0.85 (See Tables 5A and 5B). The results are less sensitive to the controls for selection. This is an interesting finding that suggests that selection might not be such a big problem on longer-term outcomes such as having graduated from college and study early labor market outcomes. When looking at Spearman Rank correlations across rankings based on labor participation and SABER

PRO, we observe that both rankings are positively correlated. However, the correlation coefficients decrease once controls for selection are added. These results suggest that within programs like Agriculture and Veterinary (with relatively lower correlations in the ranking based on SABER PRO and labor market contributions) there might be colleges that are not adding in terms of knowledge and skills but that are adding in terms of labor market outcomes.

On the other hand, within other program areas like Math and Natural Sciences (with the highest rank correlation) we find that most colleges are adding both knowledge and skills and labor market outcomes even after controlling for selection. This is a different result than the one presented above when we compared the correlation of rankings based on SABER PRO and the probability of graduation. These results clearly illustrate the need to correct for selection, as well as the need to look at relevant outcomes separately to identify how particular college-program combinations might be doing a better job at specific educational and labor outcomes. We also studied the degree of correlation across rankings based on graduation rates and labor market outcomes. In this case we observe that, although the correlations become smaller and in some cases reverse sign when controlling for selection, for the areas of Math and Natural Sciences and Social Sciences and Humanities, rankings based on these two outcomes stay positively correlated with and without controls for selection.

<<Tables 5A and 5B>>

Ranking according to college-program contributions to Initial Wages. We also study college-program contributions to beginning wages among graduates. As in the case of SABER-PRO, we follow a linear regression model on specification (1) above. The results suggest that program-college contributions to initial monthly wages range from 910 to 1,517 Colombian pesos. Every program-college has relatively large averages, and there is less variation between program-college combinations (See Tables 6A and 6B). As is the case for employment, these results are less sensitive to the controls for selection. The results of the correlations between SABER PRO and initial wages outcomes show that some correlations between those outcomes actually become negative once the controls for selection are added.

These results suggest that programs like Education and Social Sciences and Humanities might not be adding in terms of knowledge and skills but are adding in terms of early wages. On the other hand, programs like Math and Natural Sciences are adding both knowledge and skills and labor market outcomes even after controlling for selection. These results clearly illustrate the need to correct for selection, as well as the need to look at the outcomes separately, given that the same institutions that might be ranked lower according to their contribution to knowledge and skills might be ranked higher in terms of their contributions to graduation and early labor market outcomes.

<<Tables 6A and 6B>>

Robustness Check: Heckman Correction to address the problem of drop out rates As described above, an additional complication of computing value-added estimates comes from the potential bias due to differential non-random dropout rates of students and the fact that only students who graduate take the SABER PRO exam. To the extent that the demographic variables and test scores in SABER 11 are also important predictors of dropout rates, by controlling for them we would also take into account this potential bias. However, these proposed methods won't control for potential unobservables determining graduation that could also be linked to the unobservables determining SABER PRO test results. To take this into account we used the Heckman (1978) traditional selection correction models for rankings based on SABER-PRO. To help identification of these models we added information about unemployment rates in the student's area of residence at the last time we observed the student enrolled in college. This is our exclusion restriction; we assume that this variable will determine dropout decisions but will not affect the results of SABERPRO directly. In particular, our selection equation for the probability of being observed taking SABERPRO included the same socio-demographic controls as described above for the main equation in (1), as well as college effects and local unemployment rate. We then estimated two versions of the Heckman selection's model: 1) Including SABER 11 scores and cohort average SABER 11 scores to correct for selection into different colleges and programs both in the SABERPRO equation and in the selection equation, and 2) excluding the SABER 11 and cohort average SABER 11 controls from both the main and selection equations. The local unemployment rate, our exclusion restriction variable, presented negative and highly significant effects in all our estimates using Heckman's selection model. This indicates that our sample colleges located in areas with higher unemployment rates have students with a higher probability of dropping out. We found that sample selection due to students dropping out from college was an issue when controls for SABER 11 and cohort average SABER 11 were included in the model. The estimated Heckman's lambda coefficient was 0.07 and significant at the 99% level in this case. This indicates a positive selection in our sample. We did not find a significant lambda coefficient when SABER 11 and cohort average SABER 11 controls were excluded from the model. We believe this could be an indication that these models were not correctly specified.[9][10]

Our results for college-program contributions to SABERPRO without and without SABER 11 and cohort average SABER 11 controls are presented in Tables 7A and 7B and Figures 2A

[9] A full set of results for the Heckman selection models is available from the authors upon request.

[10] We also estimated Heckman selection models for equations analyzing college-program effects on initial salaries, among college graduates. Selection bias could occur in this case because of the non participation of some graduates in the labor market. Our estimates, however, suggested that sample selection did not seem to be much of a problem in this case.

and 2B. Consistent with the FE estimations, once we introduce the SABER 11 and cohort average SABER 11 controls, the VA gains in SABER PRO from different program-college combinations diminish substantially. However, similarly to the FE estimates without Heckman's corrections, for institutions in the top 75th percentile of the distribution, there is evidence of positive gains in general knowledge. Comparisons between the FE models with and without the Heckman correction suggest that such corrections make a difference, as can be seen by the relatively low Spearman Rank correlations across rankings based on models with and without Heckman corrections presented in Tables 7A and 7B. Also, without such corrections we would be overestimating the contribution of postsecondary institutions in terms of knowledge and skills. These findings confirm how sensitive the ranking is to different model specifications and the problems associated with making high-stakes decisions based on these estimates.

<<Tables 7A and 7B>>

Conclusions

A number of important findings emerged from the analyses conducted in this paper. First, any college system or country attempting to develop a system of indicators to rank post-secondary education institutions on a number of relevant educational and labor market outcomes needs to be aware of the challenge of producing unbiased estimates and the need to correct for the problem of selection of students into institutions. Our findings clearly illustrate that once we addressed the selection issue, the initial gains in generic knowledge basically disappeared. Second, in this paper we not only provided estimates using three different types of models (i.e., AR, RA, and FE) but we also present correlations or rankings within these methods for different specifications. By doing this we are providing solid empirical evidence in favor of FE models that deals with the issue of student selection. Third, in the paper we also empirically tested the

Heckman correction to address the issue related to the “survivor bias” introduced by having only information from the students who took the SABER PRO exam. Our findings suggested that the FE models without such correction might be overestimating the college-program contributions in terms of generic knowledge and rankings should be created using these two complementary methods to correct for both selection into certain colleges and programs and selection due to students dropping out from college. Finally, our findings confirm our hypotheses that rankings of specific college-program combinations change depending on different educational and labor outcome measures considered. This is a very important finding that emphasizes the need to use many complementary indicators related to the mission of the specific post-secondary institutions that are being ranked.

Even though the main objective of this study was to contribute to a more nuanced understanding of the methods that need to be used to minimize bias as systems attempts to rank institutions, our findings also contribute to the growing literature related to measuring SLOs in higher education. The overall findings suggest that the majority of postsecondary institutions in Colombia, during the period of time studied, were not really contributing in terms of adding generic knowledge, above what was expected given student characteristics; instead, these institutions added value by providing the students with the diplomas and certifications necessary to enter the labor market and benefit from the economic return of their degrees. This provides compelling evidence (as hypothesized by Barrera-Osorio and Bayona-Rodriguez (2014)) that for this specific combination of college-programs, students were benefiting from the signaling provided by their degrees, rather than from the curriculum itself. This finding has competing explanations. One possible explanation is that the test that is being used to measure the generic skills is not doing a good job measuring these competencies. This might not be the case for this particular study given that the generic component was developed based on the CLA, a test that has strong validity and reliability (Steedle, 2012). An alternative explanation is that the curriculum of the majority of the programs is mostly focusing on providing the subject specific skills. Melguizo and Wainer (2014) found evidence supporting this explanation for the case in Brazil.

These findings lead us to question whether the majority of post-secondary education institutions are interested in cultivating generic skills such as critical thinking and problem solving. It is important for higher education systems to outline the knowledge and skills that they want institutions to provide, so these institutions can develop both relevant curriculum and appropriate assessment tools. This is much more complicated in higher education than in the K-12 system because of the great institutional autonomy of postsecondary institutions and the fact that public support for higher education has decreased substantially.

The results of this paper illustrate the importance of validating empirical models intended to rank college-program contributions according to a number of educational and early labor market outcomes. The results also suggest that given the sensitivity of the models to different specifications, it is not clear that they should be used to make any high-stakes decisions in higher education. They could, however, serve as part of a broader set of indicators to support programs and colleges as part of a formative evaluation.

In summary, in line with recommendations of the AHELO program, the results of this study highlight the importance for higher education systems to think about multi-faceted accountability measures that are closely aligned with the mission of the postsecondary institutions and contribute mainly to formative evaluations.

References

Arrow, K. J. (1973). Higher education as a filter. *Journal of Public Economics*, *2*(3), 193–216.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. University of Chicago Press.

Barrera-Osorio, F., & Bayona-Rodríguez, H. (2014). The causal effect of university quality on labor market outcomes: Empirical evidence from Colombia. Presented at the V Seminario Internacional ICFES sobre Investigación en la Calidad de la Educación, Bogotá, Colombia.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *The Journal of Political Economy*, 9–49.

Briggs, D. C. (2012). Making value-added inferences from large-scale assessments. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large-scale assessment in education: Theory, issues and practice* (pp. 186–206). New York, NY: Routledge.

Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics, 66*(2), 103–115.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2633–79. doi:10.1257/aer.104.9.2633

Coates, H. (2009). What’s the difference? A model for measuring the value added by higher education in Australia. *Higher Education Management and Policy*, *21*(1), 1–20.

Cunha, J. M., & Miller, T. (2014). Measuring value-added in higher education: Possibilities and limitations in the use of administrative data. *Economics of Education Review*, *42*, 64–77. doi:10.1016/j.econedurev.2014.06.001

Domingue, B.W., Morales, J.A., Shavelson, R., Wiley, E., Molina, A., & Mariño, J.P. (2014). *Challenges to the study of school effects in higher education*. Institute of Behavioral Science at the University of Colorado Boulder, Instituto Colombiano para la Evaluación de la Educación in Bogotá, Colombia, SK Partners, LLC, and Stanford University.

Field, K. (2013, August 22). Obama plan to tie student aid to college ratings draws mixed reviews. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Obama-Proposes-Tying-Federal/141229/

Gorard, S. (2008). The value-added of primary schools: what is it really measuring? *Educational Review*, *60*(2), 179–185.

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, *46*(4), 931–959. doi:10.2307/1909757

Kane, T.J., & Staiger, D.O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Working Paper No. 14607). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14607

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review*, 31(5), 415–439.

McCaffrey, D.F., Lockwood, J.R., Koretz, D.M., & Hamilton, L.S. (2003). *Evaluating value-added models for teacher accountability*. RAND Corporation.

Melguizo, T. (2008). Quality matters: Assessing the impact of attending more selective institutions on college completion rates of minorities. *Research in Higher Education, 49*(3), 214–236. doi:10.1007/s11162-007-9076-1

Melguizo, T., & Wainer, J. (2014). Are students gaining general and subject area knowledge in university? Evidence from Brazil. Presented at the V Seminario Internacional ICFES sobre Investigación en la Calidad de la Educación, Bogotá, Colombia.

Organization for Economic Co-operation and Development. (2013). *Education at a glance 2013*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/eag-2013-en

Organization for Economic Co-operation and Development. (2014, June). Testing student and university performance globally: OECD’s AHELO. Retrieved February 14, 2015, from www.oecd.org/edu/ahelo

Pérez-Peña, R. (2013, October 8). U.S. adults fare poorly in a study of skills. *The New York Times*. Retrieved from http://www.nytimes.com/2013/10/08/us/us-adults-fare-poorly-in-a-study-of-skills.html

Saavedra, A. R., & Saavedra, J. E. (2011). Do colleges cultivate critical thinking, problem solving, writing and interpersonal skills? *Economics of Education Review*, *30*(6), 1516–1526. doi:10.1016/j.econedurev.2011.08.006

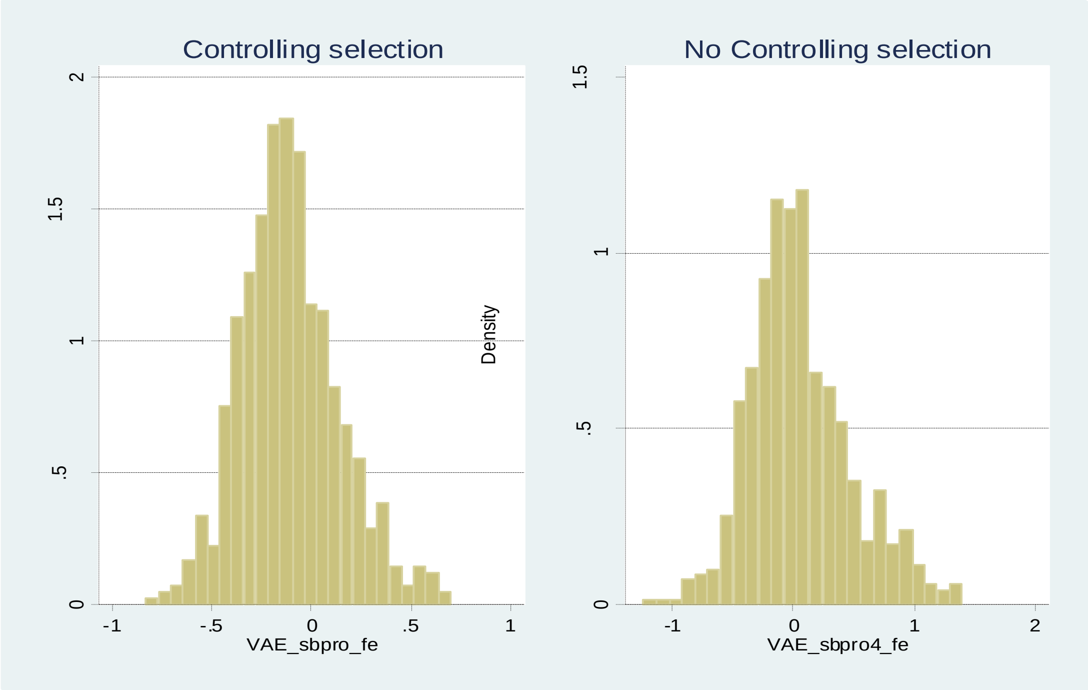
Saavedra, J. E. (2009). *The learning and early labor market returns to college quality: A regression discontinuity analysis*. Harvard University, Cambridge, MA.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2012). Value added of teachers in high-poverty schools and lower poverty schools. *Journal of Urban Economics*, *72*(2-3), 104–122. doi:10.1016/j.jue.2012.04.004

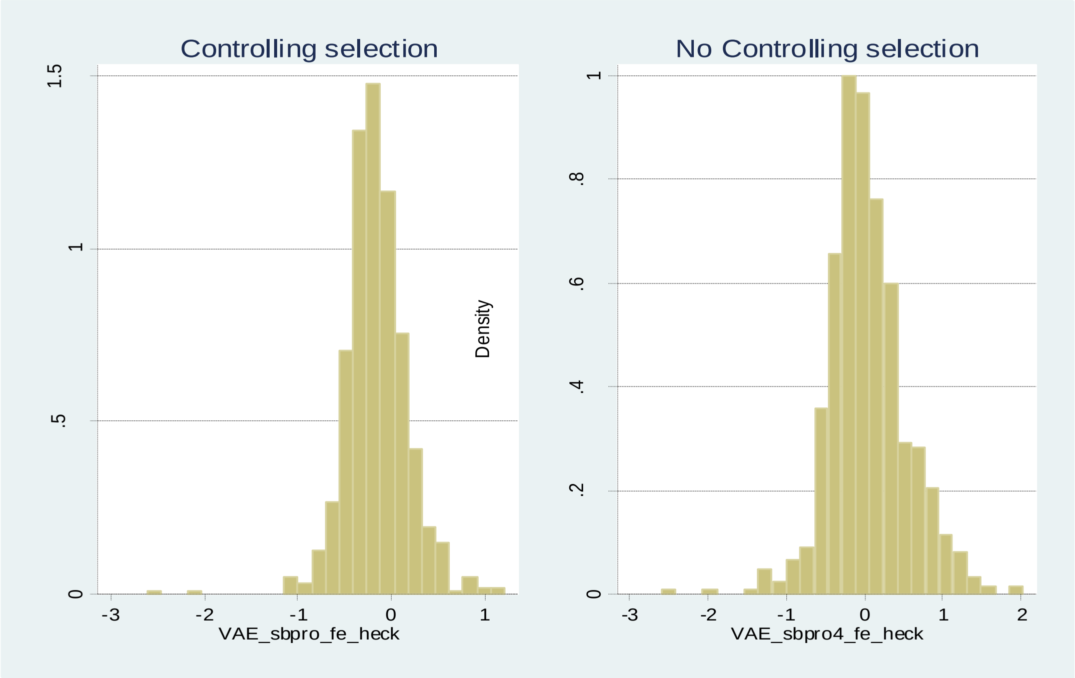
Spellings, M. (2006). *A test of leadership: Charting the future of US higher education* (U.S. Department of Education Contract No. ED-06-C0-0013). Jessup, MD: Education Publications Center.

Steedle, J. T. (2012). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education*, *37*(6), 637–652.

Stratford, M. (2015, January 30). Counting students equally? *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/news/2015/01/30/ed-dept-ratings-framework-ignites-new-questions-over-adjusting-student-outcomes.



Figures 1A and 1B. Value-added estimates controlling and no selection controls



Figures 2A and 2B. Value-added estimates controlling and no selection controls

Table 1. *Descriptive Statistics for Entering Cohorts 2004 Onwards With Expected Graduation Rates from 2009 to 2012*

	Average outcomes by Program-College									
	SABER PRO		SABER 11		Labor rates		Graduation rates		Wages	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Agriculture, Veterinary	0.16	0.44	0.39	0.39	0.67	0.29	0.40	0.19	931.89	246.89
Art	0.34	0.44	0.63	0.40	0.62	0.25	0.42	0.17	1044.67	504.71
Education	-0.09	0.43	0.09	0.28	0.57	0.25	0.45	0.14	908.99	181.34
Health	0.21	0.47	0.52	0.48	0.71	0.22	0.58	0.18	1523.29	562.08
Social Sciences and Humanities	0.19	0.41	0.34	0.35	0.60	0.24	0.44	0.18	1247.65	326.75
Economics and Business	0.18	0.45	0.35	0.36	0.68	0.23	0.48	0.16	1081.62	318.84
Engineering, Architecture	0.33	0.43	0.57	0.40	0.71	0.24	0.44	0.17	1231.28	574.79
Math and Natural Sciences	0.42	0.53	0.66	0.50	0.65	0.20	0.46	0.16	1204.33	475.39

Table 2A. *Spearman Rank Correlation Coefficients across Methods*

	ρ-AR vs FE	ρ-RE vs FE	ρ-AR vs RE
Controlling for Selection	0.77	0.86	0.91
Not Controlling for Selection	0.96	0.98	0.95
Controlling for SABER 11	0.96	0.95	0.92
Controlling for Peer Effects	0.77	0.85	0.92

Table 2B. *Spearman Rank Correlation Coefficients within Methods-Compared to Controlling for Selection*

	ρ-FE	ρ-RE	ρ-AR
Not Controlling for Selection	0.91	0.67	0.52
Controlling for SABER 11	0.99	0.96	0.94
Controlling for Peer Effects	0.94	0.92	0.91

Table 3A. *Distribution of Value-Added Program-College Contributions-FE Model*

	Mean	SD	25%	50%	75%	n
Agriculture, Veterinary	-0.13	0.23	-0.27	-0.15	0.01	41
Art	-0.15	0.22	-0.26	-0.14	-0.04	49
Education	-0.18	0.30	-0.38	-0.18	-0.14	75
Health	-0.16	0.23	-0.31	-0.21	-0.06	69
Social Sciences and Humanities	-0.08	0.24	-0.27	-0.10	0.06	123
Economics and Business	-0.10	0.26	-0.28	-0.14	0.04	156
Engineering, Architecture	-0.07	0.22	-0.21	-0.09	0.06	132
Math and Natural Sciences	-0.03	0.25	-0.21	-0.05	0.17	30

Table 3B. *Distribution of Value-Added Program-College Contributions-FE Model- No Selection Controls*

	Mean	SD	25%	50%	75%	n
Agriculture, Veterinary	0.02	0.40	-0.23	0.03	0.29	41
Art	0.09	0.40	-0.15	0.04	0.30	49
Education	-0.15	0.40	-0.42	-0.15	0.06	75
Health	0.09	0.46	-0.22	-0.05	0.38	69
Social Sciences and Humanities	0.03	0.37	-0.24	0.00	0.23	123
Economics and Business	0.03	0.40	-0.23	-0.07	0.24	156
Engineering, Architecture	0.15	0.40	-0.12	0.10	0.40	132
Math and Natural Sciences	0.29	0.49	0.01	0.14	0.69	30

Table 4A. *Distribution of Program-College Contributions to Graduation Rates: LPM with FE Model*

	Mean	SD	25%	50%	75%	ρ-SABER PRO Graduation	n
Agriculture, Veterinary	0.46	0.18	0.34	0.42	0.58	0.01	43
Art	0.44	0.16	0.35	0.46	0.56	-0.24	55
Education	0.57	0.15	0.44	0.56	0.68	-0.21	83
Health	0.61	0.18	0.52	0.62	0.73	-0.31	75
Social Sciences and Humanities	0.50	0.17	0.41	0.51	0.61	0.05	129
Economics and Business	0.55	0.15	0.46	0.56	0.63	0.08	154
Engineering, Architecture	0.47	0.16	0.39	0.48	0.56	-0.19	147
Math and Natural Sciences	0.47	0.15	0.37	0.44	0.55	-0.06	36

Table 4B. *Distribution of Program-College Contributions to Graduation Gates: LPM with FE Model, No Selection Controls*

	Mean	SD	25%	50%	75%	ρ-SABER PRO Graduation	n
Agriculture, Veterinary	0.49	0.19	0.34	0.47	0.60	0.31	43
Art	0.48	0.16	0.39	0.50	0.61	0.08	55
Education	0.56	0.14	0.45	0.55	0.67	-0.11	83
Health	0.65	0.18	0.58	0.65	0.78	0.19	75
Social Sciences and Humanities	0.52	0.17	0.41	0.53	0.64	0.30	129
Economics and Business	0.57	0.16	0.48	0.58	0.66	0.29	154
Engineering, Architecture	0.51	0.11	0.41	0.52	0.63	0.14	147
Math and Natural Sciences	0.54	0.16	0.42	0.54	0.61	0.25	36

Table 5A. *Distribution of Program-College Contributions to Success in Labor Market: LPM with FE Model*

						ρ-SABER	ρ-Labor	
						PRO	Market	
	Mean	SD	25%	50%	75%	LaborMarket	Graduation	n
Agriculture, Veterinary	0.81	0.29	0.63	0.86	1.01	0.09	0.05	59
Art	0.74	0.24	0.59	0.71	0.92	0.26	-0.07	71
Education	0.70	0.25	0.57	0.74	0.89	0.33	-0.58	92
Health	0.84	0.23	0.74	0.83	0.98	0.27	-0.15	87
Social Sciences and Humanities	0.72	0.23	0.57	0.74	0.87	0.39	0.15	144
Economics and Business	0.81	0.22	0.69	0.83	0.95	0.40	-0.21	179
Engineering, Architecture	0.83	0.23	0.72	0.87	0.98	0.29	-0.11	176
Math and Natural Sciences	0.79	0.21	0.64	0.81	0.92	0.47	0.26	42

Table 5B. *Distribution of Program-College Contributions to Success in Labor Market: LPM with FE Model, No Selection Controls*

						ρ-SABER	ρ-Labor	
						PRO	Market	
	Mean	SD	25%	50%	75%	Labor Market	Graduation	n
Agriculture, Veterinary	0.81	0.29	0.64	0.86	1.01	0.20	0.20	59
Art	0.74	0.24	0.60	0.72	0.93	0.39	0.09	71
Education	0.70	0.25	0.57	0.75	0.90	0.35	0.06	92
Health	0.85	0.23	0.74	0.83	0.99	0.28	-0.06	87
Social Sciences and Humanities	0.72	0.24	0.56	0.75	0.87	0.42	0.29	144
Economics and Business	0.81	0.22	0.69	0.83	0.96	0.53	-0.02	179
Engineering, Architecture	0.84	0.23	0.72	0.89	0.99	0.36	0.09	176
Math and Natural Sciences	0.80	0.21	0.65	0.82	0.93	0.46	0.39	42

Table 6A. *Distribution of Program-College Contributions to Initial Wages-LPM with FE Model*

	Mean	SD	25%	50%	75%	ρ-SABER PRO Initial Wages	n
Agriculture, Veterinary	910.36	232.76	757.42	898.47	1027.60	0.12	54
Art	1020.14	516.15	769.51	912.95	1114.22	0.26	68
Education	917.78	189.78	810.93	893.25	1010.99	-0.06	87
Health	1476.99	521.23	1164.92	1474.10	1765.03	0.55	84
Social Sciences and Humanities	1215.22	300.97	999.23	1171.66	1400.67	-0.04	136
Economics and Business	1080.67	287.49	907.26	1032.84	1208.16	0.33	176
Engineering, Architecture	1195.04	564.13	980.07	1137.85	1304.76	0.33	168
Math and Natural Sciences	1113.06	452.08	810.86	1026.78	1263.10	0.25	40

Table 6B. *Distribution of Program-College Contributions to Initial Wages: LPM with FE Model, No Selection Controls*

	Mean	SD	25%	50%	75%	ρ-SABER PRO Initial Wages	n
Agriculture, Veterinary	944.60	237.56	785.92	904.54	1039.18	0.32	54
Art	1052.97	508.80	800.63	976.06	1108.41	0.31	68
Education	935.21	197.16	827.52	909.29	1029.55	0.10	87
Health	1517.82	540.60	1172.31	1508.71	1824.13	0.64	84
Social Sciences and Humanities	1238.83	326.54	1020.88	1195.47	1430.55	0.06	136
Economics and Business	1110.98	299.03	940.37	1062.21	1241.61	0.38	176
Engineering, Architecture	1243.33	572.86	1009.92	1184.79	1366.23	0.49	168
Math and Natural Sciences	1188.46	469.28	890.55	1149.48	1316.40	0.34	40

Table 7A. *Distribution of Value-Added Program-College Contributions-- Heckman with FE*

Model

	Mean	SD	25%	50%	75%	ρ-SaberPro with and w/out Heckman	N.obs
Agriculture, Veterinary	-0.1455	0.2967	-0.3479	-0.1341	0.0304	0.5506	46
Art	-0.1065	0.2475	-0.2786	-0.1710	0.0674	0.5994	61
Education	-0.1716	0.2972	-0.3577	-0.2056	-0.0351	0.3563	75
Health	-0.1448	0.2555	-0.3050	-0.1704	0.0198	0.2298	69
Social Sciences and Humanities	-0.1629	0.3254	-0.3775	-0.1887	0.0265	0.2716	124
Economics and Business	-0.1590	0.3451	-0.3696	-0.1896	0.0338	0.2685	151
Engineering, Architecture	-0.1701	0.4006	-0.3592	-0.1893	0.0367	0.2540	132
Math and Natural Sciences	-0.1986	0.4475	-0.3025	-0.1833	-0.0208	0.5573	37

Table 7B. *Distribution of Value-Added Program-College Contributions: Heckman with FE Model, No Selection Controls*

	Mean	SD	25%	50%	75%	ρ-SaberPro with and w/out Heckman	N.obs
Agriculture, Veterinary	0.0390	0.4753	-0.2704	0.0055	0.2992	0.4890	46
Art	0.1117	0.4160	-0.1957	0.0416	0.3511	0.6406	61
Education	-0.0068	0.4506	-0.2237	-0.0459	0.1774	0.3641	75
Health	0.0882	0.4061	-0.1790	0.0267	0.3542	0.3283	68
Social Sciences and Humanities	0.0513	0.4820	-0.2699	0.0052	0.3301	0.2890	124
Economics and Business	0.0496	0.5032	-0.2562	0.0090	0.3233	0.3002	151
Engineering, Architecture	0.0154	0.5819	-0.3125	-0.0230	0.3641	0.2989	132
Math and Natural Sciences	0.0473	0.6157	-0.1797	-0.0151	0.3449	0.3281	37