Measuring Effects of Colombian Postsecondary Institutions on Student Learning

Ben Domingue
with thanks to Ed Wiley

Institute of Behavioral Sciences
University of Colorado Boulder
ben.domingue@gmail.com

ICFES: November 2012

Why Measure Learning Effects?

- ▶ Student learning is primary function of post-secondary education.
- ▶ Can we measure institutional differences in student learning?
- ▶ If so, this would be useful to a variety of stakeholders (students, parents, government, business, etc.)

How do we gauge whether institutions have influenced student learning?

Why Measure Learning Effects?

- ▶ Student learning is primary function of post-secondary education.
- ▶ Can we measure institutional differences in student learning?
- ▶ If so, this would be useful to a variety of stakeholders (students, parents, government, business, etc.)

How do we gauge whether institutions have influenced student learning?

- ▶ Ideally, what we'd like is to have a measure of student achievement before *and* after they receive post-secondary education.
- ▶ A *standardized* measure of ability at the end of post-secondary education is more or less non-existent in most countries.

Colombia is Unique

- Colombia has a unique opportunity as the same students take SABER 11 and SABER PRO.
- Being in this unique position makes this research particularly compelling.

Colombia is Unique

- Colombia has a unique opportunity as the same students take SABER 11 and SABER PRO.
- Being in this unique position makes this research particularly compelling.

In Colombia, an initial study on the effectiveness of post-secondary education has already been done. This serves as a good starting point for discussing this project.

Saavedra & Saavedra Findings

Important: Interest was on the effect of post-secondary education, not the effect of individual institutions of post-secondary education.

- ▶ Saavedra & Saavedra used data from the Graduate Skills Assessment Test (ACER's MC test translated into Spanish) for roughly 2,000 students finishing college (matched to entering college students) from a sample of 17 institutions (representative of the top 100 universities nation-wide) to estimate "how much value colleges add to students' critical thinking, problem-solving, and communication skills".

Saavedra & Saavedra Findings

Important: Interest was on the effect of post-secondary education, not the effect of individual institutions of post-secondary education.

- ▶ Saavedra & Saavedra used data from the Graduate Skills Assessment Test (ACER's MC test translated into Spanish) for roughly 2,000 students finishing college (matched to entering college students) from a sample of 17 institutions (representative of the top 100 universities nation-wide) to estimate "how much value colleges add to students' critical thinking, problem-solving, and communication skills".

- ▶ On the overall test, completing university is associated with approximately a 0.5 SD increase on the GSA measure.

- ▶ This is smaller than the 1 SD effect found for US universities (Klein et al., 2007).

Saavedra & Saavedra Findings

Additional questions:

- Are private and public institutions differentially effective?
- Are metrics of quality (selectivity, % of faculty with PhD) associated with effectiveness?

Saavedra & Saavedra Findings

Additional questions:

- ▶ Are private and public institutions differentially effective?
- ▶ Are metrics of quality (selectivity, % of faculty with PhD) associated with effectiveness?

They found:

- ▶ Private institutions were more effective than public institutions (0.31 SDs higher).
- ▶ Typical metrics of quality (such as selectivity or % of faculty with a PhD) are not especially predictive of effectiveness. (Note: This echoes findings from studies American teacher effectiveness that the qualities that teachers typically receive additional money for—experience, additional education (e.g., masters degree in curriculum)—don't translate into additional effectiveness.

Saavedra & Saavedra Findings

They also found that universities are differentially effective, an empirical motivation for this work.

Shortcomings of Saavedra & Saavedra

- Saavedra & Saavedra used cross-sectional data (students who were entering college were contrasted with students who were leaving college). This can lead to biased effect estimates for several reasons.
- Primarily because the cohorts of students may have differed.

Shortcomings of Saavedra & Saavedra

- ▶ Saavedra & Saavedra used cross-sectional data (students who were entering college were contrasted with students who were leaving college). This can lead to biased effect estimates for several reasons.
- ▶ Primarily because the cohorts of students may have differed.
- ▶ Recall that in the "ivory tower" ideal we'd measure changes in achievement for *every student who entered college*, linking each student's scores upon entry to and exit from college.
- ▶ Available longitudinal data is not perfect, but it is potentially superior then stricly cross-sectional analyses.
- ▶ New study will examine university effectiveness using longitudinal data. Focus will be on students who took the SABER PRO in 2011.

Test Scores

Since test scores are at the heart of this endeavor, let's take a closer look.

- A student's college exit score (the *outcome*) will be a student's score from the 2011 administration of the SABER PRO.
- A student's college entry score (the *prior*) will be a student's score from when they took the SABER 11.

Test Scores

This study relies upon linked SABER PRO and 11 scores. Data from the survey administered to students when they took the SABER PRO also used.

SABER PRO—2011 Administration

- Reading, Writing, Quantitative & English
- SD of 1.

SABER 11—Whenever a student from the 2011 SABER PRO group took it

- Math, Language, & English
- Standardized within each administration.

Test Scores

This study relies upon linked SABER PRO and 11 scores. Data from the survey administered to students when they took the SABER PRO also used.

SABER PRO—2011 Administration

- ▶ Reading, Writing, Quantitative & English
- ▶ SD of 1.

SABER 11—Whenever a student from the 2011 SABER PRO group took it

- ▶ Math, Language, & English
- ▶ Standardized within each administration.

Given the multitude of tests, how do we decide which SABER 11 outcome to treat as a pre-test for a SABER PRO score?

Correlation between scores

One way of making this decision is to use correlations between test scores.

		SABER PRO			
		Reading	Writing	Quant	English
	Math	0.38	0.20	**0.48**	0.43
SABER 11	Language	**0.56**	**0.32**	0.45	0.49
	English	0.46	0.29	0.44	**0.69**

Correlation between scores

One way of making this decision is to use correlations between test scores.

		SABER PRO			
		Reading	Writing	Quant	English
	Math	0.38	0.20	**0.48**	0.43
SABER 11	Language	**0.56**	**0.32**	0.45	0.49
	English	0.46	0.29	0.44	**0.69**

All but Language-Writing have relatively high correlations.

- The English test wasn't always required on the SABER 11.
- Because its correlation is higher than the Math-Quant correlation, this study focuses on the Language-Reading sequence.

What do we need?

- ▶ Remember our goal: to *fairly* measure university effectiveness using longitudinal data
- ▶ What we need: a measure of variation in effectiveness among institutions that creates a level playing field with respect to differences in incoming performance and student characteristics.
- ▶ Statistical models that attempt to measure the effectiveness of institutions as we've described here are known as *value-added models*.

Value-Added (VA) Models

A value-added model represents a statistical approach that attempts to disentangle the effect of institutions from other factors (e.g., parent wealth, student motivation) that contribute to student learning. Although there are many different ways to specify a value-added model, they all have two things in common:

Value-Added (VA) Models

A value-added model represents a statistical approach that attempts to disentangle the effect of institutions from other factors (e.g., parent wealth, student motivation) that contribute to student learning. Although there are many different ways to specify a value-added model, they all have two things in common:

- ▶ All value-added models control for a student's incoming achievement level. This is done to ensure that an institution is not unfairly held accountable for preexisting differences among students. The focus of a value-added model is on conditional achievement or achievement *growth*.

Value-Added (VA) Models

A value-added model represents a statistical approach that attempts to disentangle the effect of institutions from other factors (e.g., parent wealth, student motivation) that contribute to student learning. Although there are many different ways to specify a value-added model, they all have two things in common:

- ▶ All value-added models control for a student's incoming achievement level. This is done to ensure that an institution is not unfairly held accountable for preexisting differences among students. The focus of a value-added model is on conditional achievement or achievement *growth*.

- ▶ No value-added model can reveal why some institutions appear more effective than others. Understanding the differences in educational practices that cause students to learn more in certain settings requires additional targeted research.

Reasoning behind VA Work

- ▶ Our goal here is to use a statistical model to (1) measure the "learning" of individual students, then (2) assess whether students at some universities learned more than others.
- ▶ We could just take each student's SABER PRO score, and subtract their SABER 11 score, and then average that for each university. However, this might be insufficient to account for differences between institutions in the types of students they enroll. For example, some institutions enroll students with high SABER 11 scores and some institutions enroll students with low SABER 11 scores.

Reasoning behind VA Work

- ▶ Furthermore, institutions have very different demographic mixes of students. We'll see, for example, that some institutions' median student is in the first economic strata while other institutions' median student is in the fifth.
- ▶ How do we make fair and sensible comparisons across such different groups?

Reasoning behind VA Work

▶ Furthermore, institutions have very different demographic mixes of students. We'll see, for example, that some institutions' median student is in the first economic strata while other institutions' median student is in the fifth.

▶ How do we make fair and sensible comparisons across such different groups?

▶ This is where value added models come in. We construct a value added model in a way that attempts to adjust for differences in the groups of students that attend different universities: differences in SABER 11 scores, and/or differences in student demographic characteristics such as economic status or maternal education.

Questions in constructing a VA model

- ▶ What if you had scores on both SABER 9 and SABER 11. Should you use them both?
- ▶ Should you adjust for student population differences using every single student demographic measure you have available? Or just a subset? Are some variables more important to consider than others?
- ▶ How do you deal with missing data? Should you just exclude students without full data?
- ▶ The majority of students complete the SABER PRO in their 9th or 10th semester after enrolling; some students, however, take it as early as their first, second, or third semester. Should this difference be addressed by the value added model?

Lots of questions

- The primary challenge in doing value added analyses, then, is that every decision made along the way could result in changes in some universities' estimated effectiveness. In most cases there is no single obvious combination of decisions generally accepted by the majority of statisticians and policymakers.
- Hence, developing a VAM requires a great deal of forethought, research, and consideration of alternatives.

Lots of questions

- ▶ The primary challenge in doing value added analyses, then, is that every decision made along the way could result in changes in some universities' estimated effectiveness. In most cases there is no single obvious combination of decisions generally accepted by the majority of statisticians and policymakers.
- ▶ Hence, developing a VAM requires a great deal of forethought, research, and consideration of alternatives.

At this point, we could have a technical discussion about selecting VA models. However, it's more important to demonstrate the kinds of results one gets from a VA model. Hence, the rest of this presentation will use a single model as a starting point for discussions about what VA models can (and can't) do.

Demographics

Since controlling for demographics differences will be important in making reasonable comparisons, it's important to understand the demographic variables I'll be using.

> **trabaja** Indicator of whether a student worked.
>
> **estrato** Socioeconomic indicator.
>
> **semestre_cursa** Semester of university in which a student took SABER PRO.
>
> **educa_madre** Maternal education level.
>
> **year11** Year in which student took SABER 11.

Demographics

Since controlling for demographics differences will be important in making reasonable comparisons, it's important to understand the demographic variables I'll be using.

> **trabaja** Indicator of whether a student worked.
>
> **estrato** Socioeconomic indicator.
>
> **semestre_cursa** Semester of university in which a student took SABER PRO.
>
> **educa_madre** Maternal education level.
>
> **year11** Year in which student took SABER 11.

Inclusion of demographic variables amounts to allowing expectation of SABER PRO scores to differ by the demographic. For example, we may expect less growth for a student that has to work while they attend school.

Levels of estrato

1–6 Scale levels.

8 Rural students without zone classification.

Distribution of estrato
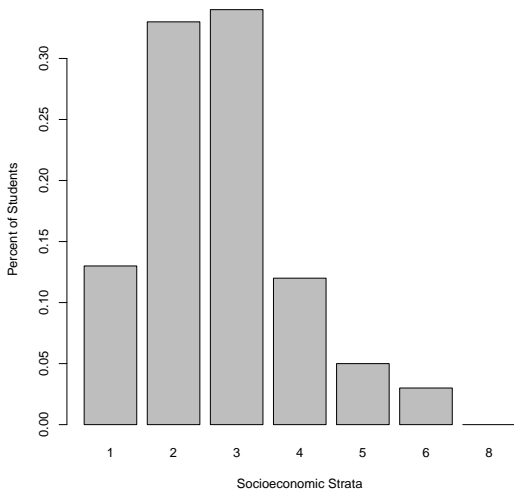
- ▶ Majority of students are in lower strata 1–3.
- ▶ Very small group of students (level 8) with no information.

Distribution of semestre_cursa
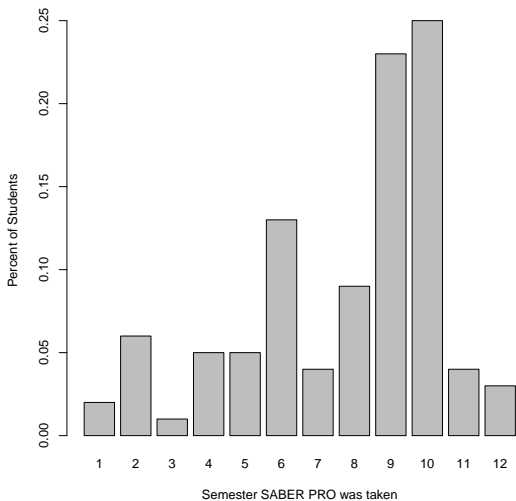
- In analysis described here, different institution types are analyzed together.

Levels of educa_madre

 0 None
 9 Did not complete primary
10 Complete primary
11 Did not complete secondary
12 Completed secondary
13 Technical education, no degree
14 Technical education with degree
15 Professional education, no degree
16 Professional education with degree
17 Postgraduate
99 Unknown

Distribution of educa_madre
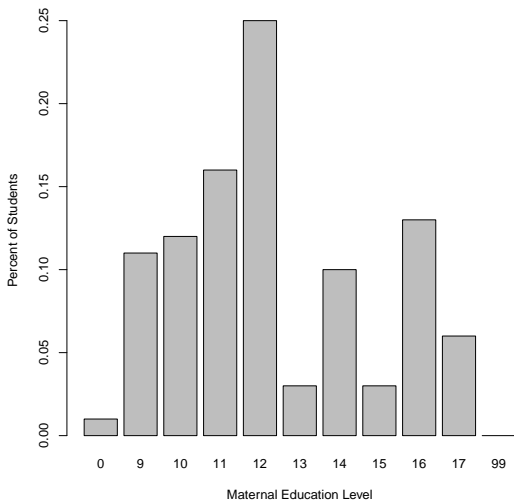
- Peaks at 12, 14, and 16 correspond to completion of high school, trade school, and university respectively.

Demographics at Institutional level

It's important to remember that note only do students vary in terms of demographics, but institutions have demographically very different types of students.
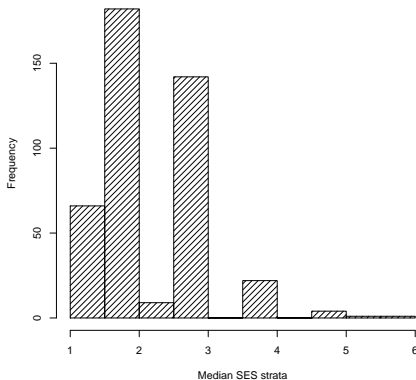
Institution Effects

Big Question: How much do institutional effectiveness estimates vary?

Variability in Effectiveness estimates
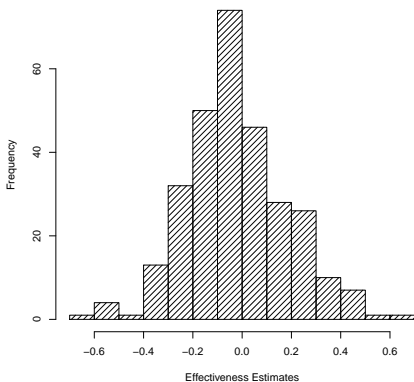
Institutions vary in the learning effects they produce.

- ► "Effectiveness estimates" are measured relative to the *average* institutions.
- ► A negative estimate, then, does not mean that an institution *removed* what had been learned at taking SABER 11, but simply reflects that that institution was less effective than the average institution.

Effectiveness in context

Now that we know institutional effectiveness varies, let's take a long look at what properties the effects have, especially as they relate to other characteristics of these institutions.

Growth versus SABER 11

SABER 11 Mean
Correlation 0.75
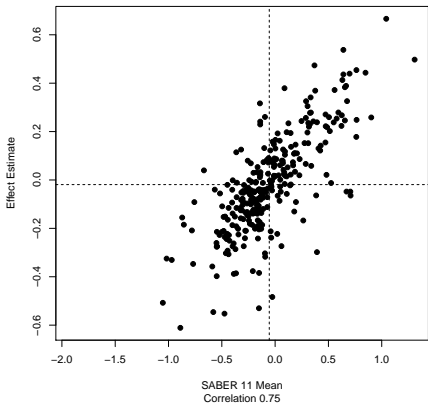
- As a reference, institution-level means between PRO and 11 were correlated at 0.79.
- Goal: "leveled the playing field" by breaking the correlation between entering student scores and their exit scores.
- Correlation here is stronger than what is commonly observed among American primary schools.

Growth versus Economic strata

SES Strata Mean
Correlation 0.46
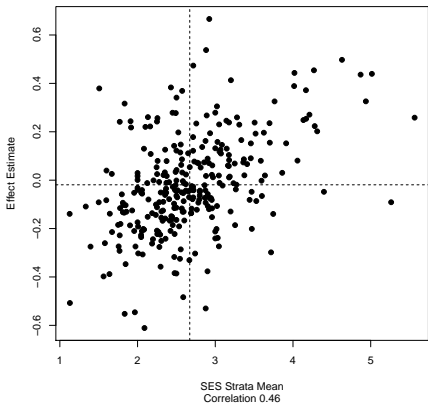
- ▶ As expected, we see a positive correlation between institutional average student socioeconomic stratum and effectiveness estimate.
- ▶ That said, throughout the socioeconomic composition spectrum, we still see substantial variation in effectiveness among institutions that share similar levels of socioeconomic composition.

VA with different units

In the same way that VA models can be used to estimate the effectiveness of post-secondary institutions, they can also be used to estimate the effectiveness of other aspects of tertiary education. Such as:

- Course of Study: 1,411 programs of study (e.g., different engineering disciplines, history, mathematics)
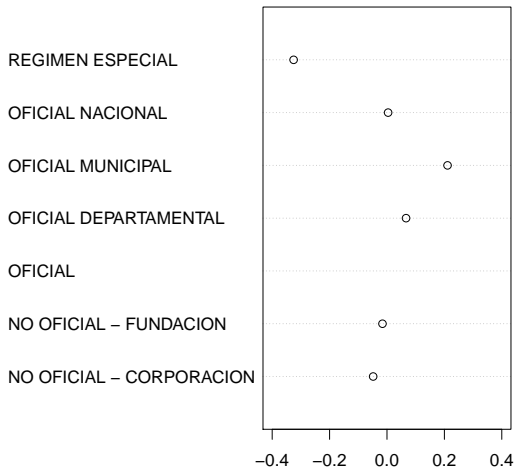- Institutional Classification
- Accreditation

Course of study
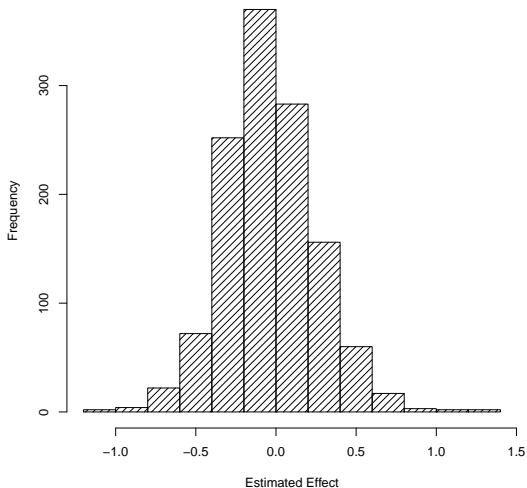
- Substantial variation for programs of study (SD of nearly 0.3 compared to 0.2 for institutions).
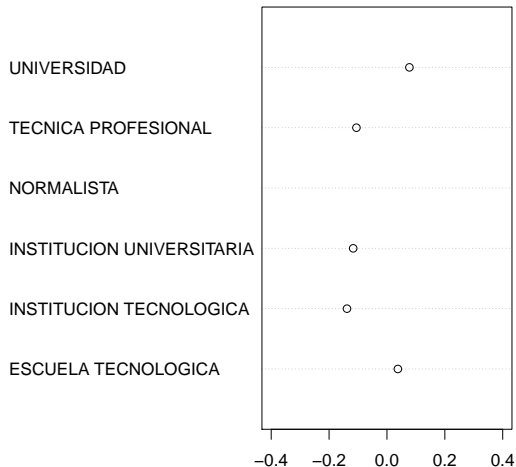
Public/Private

EB_mod_lectura_critica

- ► No estimate for "oficial" institution since there was only a single such institution.

Academic Character

EB_mod_lectura_critica

- As in previous slide, only a single "normalista" institution.

Accreditation

Accredited institutions were estimated as being 0.3 units more effective than unaccredited institutions.

- Relative to the overall student SD of 1 on this test, this is fairly sizable.

Issues to ponder

- The proliferation of results that would result if we tried to analyze all SABER PRO outcomes separately across multiple aggregations (university, area of study, etc.) would be hard to interpret. Perhaps beneficial to consider a composite.
- Considering the range of semesters over which students have completed before taking the SABER PRO, analyses could be divided into traditional universities and vocational/trade institutions.
- Model could be expanded to include additional variables, the SABER 11 mean being perhaps the most important.

Issues to ponder

- Students dropout at different rates from different universities. More concerningly, different types of students dropout from different universities. This could potentially bias estimates.

- Any statistical estimate contains uncertainty. The estimates of effectiveness can be given "confidence intervals" that suggest the variability contained in these estimates. This is especially important when discussing individual estimates.

- Measurement error on the SABER 11 test could potentially lead to biased estimates. There are methods that can correct for this that may be worth exploring.

Closing Thoughts

► Colombia is in a unique global position to do this type of research.

► One potential use of effectiveness estimates would be to offer something like a "effect per dollar" metric for all institutions in the nation. Even with the caveats that would be necessary, this would be an important indicator for stakeholders.

Thanks!

Feel free to contact me with any comments or questions.

`ben.domingue@gmail.com`