High Educ (2016) 72:381–401
DOI 10.1007/s10734-015-9963-x

Toward a set of measures of student learning outcomes in higher education: evidence from Brazil

Tatiana Melguizo[1] · **Jacques Wainer**[2]

Published online: 23 November 2015

Abstract The main objective of this study was to work toward the development of a number of measures of student learning outcomes (SLOs) in higher education. Specifically, we used data from *Exame Nacional de Desempenho dos Estudantes* (ENADE), a college-exit examination developed and used in Brazil. The fact that Brazil administered the ENADE to both freshmen and senior students provided a unique opportunity to get a first approximation of the general and subject area knowledge gained in different programs. The results suggested that, on average, students in the three different categories of programs were gaining valuable general and subject area knowledge. The gains in the subject area were of a larger magnitude than those in the general knowledge component of the test. This study contributes to the field by providing empirical and visually compelling evidence related to SLOs gains in higher education.

Keywords Student learning outcomes · Outcomes · Quality of higher education · Brazil

Introduction

Over the past five decades, higher education systems around the world have been expanding, leading to what is often referred to as the "massification" of higher education (OECD 2008). It has been argued that the broad expansion has come with a cost

✉ Tatiana Melguizo
melguizo@usc.edu

Jacques Wainer
wainer@ic.unicamp.br

[1] Rossier School of Education, University of Southern California, 3470 Trousdale Parkway, WPH 702 G, Los Angeles, CA 90089, USA

[2] Computing Institute, University of Campinas, Av. Albert Einstein 1251, Campinas, SP 13083-852, Brazil

in terms of the quality of education provided. As a result, in the last decade a number of academics, government agencies, and international organizations have started to advocate for more accountability toward measuring student learning outcomes (SLOs) in higher education (OECD 2013; Zemsky et al. 2005). The quality of higher education is a complex, contested, and elusive concept.[1] However, in a time when school budgets for higher education have been decreasing and institutions need to be more accountable to the public, it is important to start developing a set of measures related to student learning outcomes in higher education, so higher education systems can demonstrate whether students are indeed gaining valuable knowledge and skills (Coates 2009, 2014).

As a result of pressures for more accountability, in terms of outcomes provided by postsecondary institutions, a number of countries from around the world, such as England, Brazil, and Colombia, have started to create more comprehensive systems to assess the quality of higher education processes.[2] Most recently, the Organization for Economic Cooperation and Development (OECD), through its Assessment of Higher Education Learning Outcomes (AHELO)[3] project, engaged in a more comprehensive process of assessing the quality of education that includes measuring whether students are gaining appropriate general knowledge and subject area skills.

One of the countries that have been working for almost two decades to create a more comprehensive system to assess and evaluate higher education institutions is Brazil. In 1996, the Brazilian government passed the decree 2.026/96 that laid down two types of measures for assessment: (1) an analysis of general performance indicators by a number of characteristics such as state and region, area of knowledge/major field of study, and type of higher education institution and (2) an institutional assessment by peers to assess their administration, education, social integration, technological, cultural, and scientific products. This was later institutionalized in 2004 when Brazil passed federal law 10.861/2004, adopting a formal evaluation system: the Brazilian Higher Education Evaluation System or *Sistema Nacional de Avaliação da Educação Superior* (SINAES).[4] As part of SINAES, the country adopted the National Student Performance Exam or *Exame Nacional de Desempenho dos Estudantes* (ENADE),[5] which is a compulsory college-exit examination designed to measure general and subject area knowledge of students in different major fields of study (e.g., economics, engineering). Brazil has been administering the ENADE annually to students in their senior year since 2004. From 2004 through 2010, the ENADE has been administered to both senior and freshmen students in different major fields of study.

[1] We use Barnett's (1992) definition of quality of higher education (see below for a more detailed description).

[2] For a comprehensive review of countries around the world that have engaged in developing a system to evaluate SLOs in higher education, please see (Nusche 2008).

[3] The main goal of AHELO is to create a multi-dimensional, interdisciplinary, 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: http://www.oecd.org/education/skills-beyond-school/testingstudentanduniversityperformancegloballyoecdsahelo.htm.

[4] For a more detailed explanation of SINAES, please see (INEP 2009; Pedrosa et al. 2013; Verhine and Dantas 2005; Verhine et al. 2006).

[5] It is important to clarify that since 1996 the legislation established a learning outcomes test called "Exame Nacional de Cursos" that preceded ENADE.

The purpose of this study is to capitalize on the ENADE to provide some initial descriptive measures of gains in terms of SLOs in the *general* and *subject area* knowledge[6] in different major fields of study. Two main research questions guide this study: (1) Are there any gains in *SLOs*, in terms of the general and subject area scores between freshmen and seniors by specific *major field of study*? and (2) Are there any differences in *SLOs*, in the scores of the freshmen and seniors by specific *major field of study* according to specific individual (i.e., proportion of low- and high-income students enrolled in the programs) and institutional (i.e., public vs. private) characteristics?

The fact that Brazil administered the same instrument (i.e., ENADE) to both freshmen and senior students provided a unique opportunity to get a first approximation of the general and subject area knowledge and skills gained by students enrolled in different majors.

This study contributes to the field by proposing a simple, intuitive, and visually compelling methodology to present gains in SLOs in higher education. We use this methodology to estimate SLOs by program, and in addition, compare estimates by specific student and individual characteristics. We also capitalize on meta-analytic techniques to aggregate the results and present differences by specific individual (i.e., by SES) or institutional (i.e., public vs. private) characteristics. We propose a simple methodology that can be used by college administrators and state and federal officials around the world to measure SLOs in higher education. Specifically, this study estimates *effect sizes* in both the general knowledge and subject area sections of the examination. It is worth noting that this is a descriptive study and we are not addressing two issues that might be biasing the estimates. First, we are not controlling for previous academic preparation and selection of students into college and into specific programs. Second, we are only partially addressing the problem of non-random attrition. We discuss these problems in more detail in the methodology section below.

Conceptual framework and literature review

Clark (1983) characterized higher education as a triangle in which three main forces: the academic community, the state, and the market, were in constant struggle trying to shape the system toward their own particular set of beliefs and goals. The differences in their beliefs and goals for the system also implied that they have somewhat different definitions of "quality" in higher education and how to best measure and assess it. Barnett (1992) used Clark's triangle to illustrate how each of these forces shaped the current debate over the quality of higher education, and how each of these groups supports different methods to assess and evaluate higher education institutions. He concluded that the "debate over quality in higher education should be seen for what it is: a power struggle where the use of terms reflects a jockeying for position in the attempt to impose their own definitions of higher education" (Barnett 1992, p. 6).

Barnett (1992) argued that higher education is a *process* and that simple ways of measuring quality such as developing rankings of higher education institutions according

[6] The ENADE measures general knowledge which is common to all the programs participating in the study, but its content is unrelated to the student's program of study. The examination content is related to cultural and social aspects of contemporary society. The subject area includes assessments of basic areas in the undergraduate programs (http://portal.mec.gov.br/index.php?Itemid=313; Pedrosa et al. 2013). We describe the ENADE in greater detail below.

to specific outcomes (i.e., degrees awarded or employment of graduates), ignored the core mission of education: to engage in a process to help students develop the art of critical thinking and problem solving. He offered a definition of "quality" that is the one that we adopted for this study: "a high evaluation accorded to an educative process, where it has been demonstrated that, through the process, the students' educational development has been enhanced: not only have they achieved the particular objectives set for the course but, in doing so, they have also fulfilled the general educational aims of autonomy, of the ability to participate in reasoned discourse, of critical self-evaluation, and of coming to a proper awareness of the ultimate contingency of all thought and action."

The field of assessment and evaluation of higher education institutions have identified a number of important dimensions and best practices to conduct evaluations of particular programs or institutions (National Academy for Academic Leadership 2014), but to the best of our knowledge, there were no comprehensive models developed to measure student learning outcomes (SLOs) in higher education. A notable exception was a model proposed recently by Coates (2009) to measure the value-added by higher education institutions in Australia. He argued that measuring *learning* at the higher education level was a very complex issue, but it was vital for demonstrating the "quality" and "value" provided by higher education institutions. Coates listed and described four different approaches that can be combined into a single model and used to measure the "quality" and "value-added" by higher education. These four approaches are: (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, all of which could provide an independent perspective on the quality of the education provided. This study focused on the second component of Coates's model, the comparison of outcomes between objective assessments administered to cohorts in their first and later years of study, and it took advantage of having student-level scores for three different cohorts of freshmen and seniors enrolled in nineteen different undergraduate programs.

In their book "Academically Adrift," Arum and Roksa (2011) attempted to measure whether students in the USA were learning valuable skills in higher education. They used the Collegiate Learning Assessment[7] (CLA) instrument to test over 2000 freshmen in 24 institutions. The authors concluded that 45 % of students "did not demonstrate any significant improvement in learning" during the first 2 years of college. The main problem as recognized by Arum and Roksa (2011) is that the instrument lacks construct validity[8] and can only measure general skills, when the reality is that students go to college to gain subject area content. In addition, the estimates of the study might be biased because of inappropriate controls for the student's previous academic preparation and lack of controls for the problem of attrition.

A number of studies have recently attempted to measure SLOs in terms of the critical thinking and problem-solving skills gained by students in college while addressing the selection problem implicit in these types of models (Barrera-Osorio and Bayona-Rodríguez 2014; Domingue et al. 2014; Rossefsky-Saavedra and Saavedra 2011; Saavedra 2009;

[7] For a description of the Collegiate Learning Assessment (CLA), see: Klein et al. (2007).

[8] An additional problem is that to the best of our knowledge, there has not been independent evaluation of the psychometric properties of the CLA. Possin (2013), a philosophy professor, conducted a descriptive evaluation of the instrument and concluded that from the point of view of the graders of the examination, they are rendering the instrument invalid.

Steedle 2012). Rossefsky-Saavedra and Saavedra (2011) used information from two different cohorts of first-year and last-year colleges of students in 2009 in Colombia to estimate value-added models in higher education. The study used pilot data from the national postsecondary-exit examination,[9] and the final sample is composed of a selected sample of students in some major fields of study in only 17 of the 177 universities of the country. The authors estimated the value-added by institutions using regression analysis adjusting for covariates and weighed propensity scores. 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 test. A number of recent studies that have also used Colombian data present contradictory findings (Melguizo et al. 2015; Barrera-Osorio and Bayona-Rodríguez 2014; Domingue et al. 2014; Saavedra 2009). Whereas Saavedra (2009), Domingue et al. (2014) and Melguizo et al. (2015) find increases in SLOs as measured by differences between SABER 11 and SABER PRO results, Barrera-Osorio and Bayona-Rodríguez (2014) find no gains. The discrepancies on the findings might be related to differences in the model specifications, as well as the use of cohorts of students that were either part of the pilot study for SABER PRO, or who took the examination when it was not a compulsory requirement for graduation. The inconsistent findings suggest the need to explore in a systematic and rigorous way the type of models that are less subject to bias.

This study contributes to the previous literature by using the ENADE, a compulsory college-exit examination[10] to measure *both* the general knowledge and subject area skills gained by students in higher education.

National Student Performance Exam (ENADE)

The ENADE is a compulsory college-exit examination that has two main components: the general and the subject areas. The general component consists of 10 items, 8 multiple-choice (MC) questions, and 2 short essays. The subject area has 27 MC questions and 3 short essays. Students were given 4 h to complete the test. The general part was common to all programs participating in a given year, and it was unrelated to the student's program of study. It basically tested the knowledge on cultural and social aspects of contemporary society.

It is important to clarify that even though ENADE was a compulsory examination and it had high stakes for postsecondary institutions (i.e., results are used for budget allocation and accreditation), the examination is neither a prerequisite for college graduation nor a measure used by potential employers. This implies that it is not high stakes for the students and probably as a result there was substantial variation in the proportion of students who completed the examination, as well as the completion rates of the different parts of the examination. For example, in 2012 of the over 587,000 students required to take the examination, only 469,000 took the examination that year (about 80 %). In terms of the different parts of the examination, looking at the results of the ENADE 2012 in economics, about 10 % of the students did not answer any question of the multiple-choice general

[9] The Colombian Institute for the Promotion of Higher Education (ICFES-Spanish acronym) commissioned the Australian Council for Education Research (ACER) to adapt its Graduate Skills Assessment (GSA) to the Colombian university.

[10] It is important to mention that there has not been an independent psychometric study of the properties of the ENADE in all the different fields. As a result, we argue that this examination has unknown psychometric properties. To the best of our knowledge, there are only a couple of studies that have explored this issue in a rigorous way, but only in the case of the Psychology examination (Primi et al. 2010, 2011).

component, compared to about 30 % who did not answer the written questions.[11] There was also wide variation in terms of the completion rates of the examination and absenteeism by geographical regions (e.g., six geographic regions) and by control of the institution (i.e., public and private).

The ENADE had been administered annually since 2004 to students in their senior year and from 2004 to 2010 to both senior and freshmen students in different programs. At the beginning, ENADE was given to a representative sample of students, but this changed in 2009 when they changed from a sample to a census approach. The government established three main groups of programs of study: (1) Biological Sciences, (2) Science, Technology, Engineering, and Mathematics (STEM), and (3) Social Sciences, and each of these programs is evaluated every three years. In 2004, the government tested the students in 14 programs in the biological sciences (i.e., medicine and nursing), in 2005 in 20 different STEM programs (i.e., engineering and computer science), and in 2006 in 16 programs in the social sciences and business administration (i.e., sociology, economics, and business). The ENADE was administered to students enrolled in programs in each of these three categories every three years. For example, the students enrolled in the different programs of the biological science category have been assessed in 2004, 2007, 2010, and 2013. Finally, the fact that Brazil has administered the examination to both freshmen[12] and senior[13] students from 2004 until 2010 provides us with a unique opportunity to test differences by program. Finally, in terms of the reliability and validity of the ENADE only the psychology-specific examination has been evaluated (Primi et al. 2010, 2011).

Methodology

Data

We used the most recent and publicly available data for the ENADE,[14] the examination that was given to both freshmen and seniors in the three main categories of programs: (1) Science, Technology, Engineering, and Mathematics (STEM), (2) Social Sciences, and (3) Biological Sciences between 2008 and 2010. In 2008, we selected the programs of architecture, computer science, engineering,[15] physics, mathematics, and chemistry, as representative of the STEM programs. In 2009, the focus was on programs from the Social Sciences, and within this group, we selected programs in business, accounting, economics, communications, law, and tourism. Finally, in 2010 the focus was on programs from the

[11] ENADE collects information for the different parts of the examination regarding the type of participation of the student. They report: absentees, students who left the majority of the questions blank, student who protested, student who participated effectively, whether the test was not considered, whether the test was not valid, and whether there was an administrative error.

[12] Freshmen students include those students who enrolled in college the year of the ENADE and that have successfully completed at least 25 percent of the course requirements for the specific academic year.

[13] Seniors are students taking the last required courses to attain their desired degree. The examination takes place in November, and the expected graduation date is December of the year that the examination is given.

[14] All the data were downloaded from http://portal.inep.gov.br/basica-levantamentos-acessar.

[15] The ENADE divides engineering into several different subfields such as mechanical engineering, electrical engineering. We combined all these subfields into a single one: engineering. There are some methodological issues related to this choice that we explain in the Methods section and the Appendix.

Table 1 Sample sizes

Year	Total number of students taking the ENADE	Total number of programs/ departments	Total number of students included in this analysis	Total number of programs/ departments included in this analysis
2008	825,236	14,212	74,036	1851
2009	1,104,174	7460	296,978	5246
2010	650,451	4285	113,396	2944

Source: Authors' calculations using ENADE data

Biological Sciences, and within this group, we chose biology, biomedicine, physical education, nursing, pharmacy, physical therapy, medicine, nutrition, and dentistry.[16]

Sample

The sample was composed of three different cohorts of students who took the ENADE examination between 2008 and 2010. We had ENADE scores in the general and subject areas of the multiple-choice part of the examination for 484,410 students enrolled in 10,041 different programs (see Tables 1, 2). As we mentioned above, although participation in the ENADE is compulsory and it is a requirement for graduation for those selected programs of study, a student may return the answer sheet blank and would still fulfill the requirement for graduation.[17] We used a set of variables TP_PR_X that indicated whether the student effectively participated in the examination or not (we do not know how INEP determined the student's level of participation on an examination). In addition, we excluded students who did not have completed information on the list of covariates that we included in our model. Below we describe the procedure used to estimate the program-level effect sizes, the estimates from the matching models, as well as the limitations implicit in this methodological strategy.

Calculating effect sizes by program level

We first estimated the gains in SLOs of a program as the standardized difference of the mean senior scores to the mean freshmen scores of students enrolled in specific fields of study. Formally, we computed the gain as an effect size, in particular Cohen's d as:

$$d = \frac{\mu(\text{seniors}) - \mu(\text{freshmen})}{\sigma_{\text{p}}}$$

where $\mu(\text{seniors})$ was the average test score for seniors, $\mu(\text{freshmen})$ the average test score for freshmen, and σ_{p} the pooled standard deviation, which was calculated as:

$$\sigma_{\text{p}}^2 = \frac{(N(\text{seniors}) - 1)\sigma^2(\text{seniors}) + (N(\text{freshmen}) - 1)\sigma^2(\text{freshmen})}{N(\text{seniors}) + N(\text{freshmen}) - 2}$$

[16] Our rationale for removing programs within these three categories was that either they were vocational and technical programs, or that they did not have enough scores.

[17] There is evidence that in certain years groups of students decided to boycott the examination by not participating in it or not completing it (Pedrosa et al. 2013).

Table 2 Sample sizes by major field of study

Field	Number of departments	Number of freshman/seniors
STEM		
Engineering	721	17,297
Physics	96	1234
Mathematics	258	3688
Chemistry	173	3262
Computer science	471	7599
Architecture	152	3938
Social Sciences		
Economy	136	3889
Law	753	49,984
Accounting	637	17,974
Business	1356	59,807
Communication	379	14,141
Tourism	120	2026
Biological Sciences		
Nutrition	214	5104
Nursing	484	16,444
Medicine	141	8132
Physical therapy	341	7964
Pharmacy	247	7201
Dentistry	162	4928
Physical education	203	4695

Source: Authors' calculations using ENADE data

The effect size calculated for the general knowledge part of the examination for economics, for example, measured the knowledge gained in the general examination by students enrolled in any economics program in the country. For each major field of study, we computed the effect size for the multiple-choice parts of both the general and the subject area examinations. We also computed the 95 % interval of confidence for the effect sizes, based on non-centrality parameters, as implemented in the MBESS R package (Kelley 2007). Below we describe how we addressed the issue of non-random attrition of students.

Propensity score matching estimates

As described above, even though both freshman and seniors are randomly selected to take the ENADE examination, there was a problem of non-random attrition (Rossefsky-Saavedra and Saavedra 2011). The fact that the less academically prepared and less motivated students might be more likely to drop out of the program before their senior year implied that both the observed and unobserved characteristics of the freshmen and seniors were different. The non-random attrition of students is problematic, and it might result in overestimated effect sizes or estimators with an upward bias. One way to address this issue was to control dropout rates of students. According to Silva Filho et al. (2007), the annual

dropout rate of students from public universities in Brazil was around 12 %, while the dropout rate for private universities was 27 %. In order to address this problem, we used a propensity score matching (PSM) technique (Stuart 2010) to identify for each individual in the treatment group (seniors) a "similar" individual in the control group (freshman) based on a distance measure called "propensity." Only matched seniors and freshmen were used in the calculation of the effect size, and so it was likely that the freshmen were "similar" to the seniors (when they were freshmen themselves). This procedure addressed to some extent the selection bias problem.

The propensity was the probability of a student becoming a senior, given a number of covariates, that is:

$$e_i = P(\text{senior}_i/X_i) - \frac{\exp(\beta X_i)}{1 - \exp(\beta X_i)}$$

where X_i was the vector of covariates for student i, and senior_i was an indicator variable on whether student i was a senior or not. The formula above calculated the propensity score for a student as the logistic regression of being a senior given the covariates.

The distance between two students was the absolute value of the differences of the logit of the propensity score of each student, and the matching was performed at the program level—that is, for each program we selected the seniors and the most "similar" freshmen.

$$d_{ij} = \left|\log it(e_i) - \log it\left(e_j\right)\right|$$

We used as covariates a number of variables that in the literature were associated with student persistence and attainment (Melguizo 2011). The following variables were included: student's family income,[18] education level of the father and mother (questions 13 and 14 in all years), student's income and relation to family regarding support (question 6 for 2009 and 2010, question 9 in 2008), gender, student's self-declared race (questions 2 in 2009 and 2010, and 5 in 2008), student's high school (private or public) (question 17 all years), and student's type of high school diploma (question 18 all years). The propensity score matching used in this research was "nearest neighbor," that is each senior student was matched with an unmatched freshman with the closest distance (as described above), as implemented in the MatchIt R package (Ho et al. 2011). A main limitation of the PSM strategy was that we were unable to control for a number of variables that in the literature were correlated with students' persistence and attainment (i.e., previous academic preparation and other non-cognitive factors such as motivation). Despite these problems, the estimates provided by this method would provide probably an upper bound of the real effect size.

Limitations

We would like to acknowledge a number of methodological issues implicit in these types of study. First, although Brazil has a compulsory high school-exit examination (ENEM) and data from the examinations are available, there is no identifying information in either dataset that allows one to link a student's ENADE and ENEM scores, and thus we could

[18] This variable was created using the income variable that was part of the student questionnaire (i.e., question 5 in 2009, 2010, and question 7 in 2008). This variable defines income according to minimum wages of all adults living in a household. Individuals from families with total income from 0 to 3 minimum wages per month (included) were defined as low income and those with family income above 10 minimum wages high income.

not include a variable to control for previous academic preparation of the students. This is problematic given that this is a critical covariate to include to address the problem of selection of students into programs and institutions. We tried to ameliorate the problem by using matching techniques, but this is not enough and estimates should be considered descriptive and probably suffering from an upward bias. Second, even though the examination was compulsory and was a prerequisite to receive the degree, about 20 % of the students did not take it and there were differences in response rates in various parts of the examination. In addition, students in certain programs and regions of the country were protesting the examination, so their results could not be included in the analyses. This is problematic and is probably biasing the estimates. Third, with the exception of Primi et al. (2010; 2011) who conducted a psychometric evaluation of the psychology examination, there has not been an independent evaluation of the ENADE. As a result, the psychometric properties of the test are unknown. Finally, similar to the threats to validity of the findings described in the study of Rossefsky-Saavedra and Saavedra (2011), our study is also subject to maturation bias.

Results

Gains in SLOs in general and subject area by program/major field of study

We present the results of the gains in SLOs for both the general and subject area tests by three main categories of programs: (1) STEM, (2) Social Sciences, and (3) Biological Sciences. The results in Fig. 1a show the effect size gain for the freshmen and seniors who were enrolled in a STEM program in any university in the country, in both the general and subject area components of the test. The central dot for each major field of study represents the effect size of the gain for all students in that field. The horizontal line, with the whiskers, represents the 95 % confidence interval on that measure.[19] The results suggest that there were gains for all the students in terms of *general knowledge,* ranging from 0.1 to 0.2 of a standard deviation. It was noteworthy that students in physics and computer science presented the larger gains. In terms of the *subject area* component, there were also gains of a much larger magnitude ranging from 0.5 to 1 standard deviation. The programs in which students presented the larger gains were physics and architecture.

We also found gains in both general and subject area components of the test for students enrolled in programs in the Social Sciences (Fig. 1b). The gains in the *general knowledge* component ranged from about 0.05 to 0.2, with students in business administration gaining more. In terms of the *subject area* component, there were larger gains ranging from 0.4 to 0.6, in which students in accounting exhibited the most gains and a very small standard deviation in the estimates was observed.

Finally, for students enrolled in Biological Sciences programs and evaluated in 2010, one can observe the larger gains in both the general and subject area components compared to the students enrolled in the STEM and Social Sciences (Fig. 1c). For these students, the gains in the *general knowledge* component of the examination ranged from zero in medicine to 0.3 in pharmacy. In terms of the *subject area* component, the results ranged from

[19] The data used to estimate this and the other figures in the paper are summarized in Table 3 in the Appendix.

0.5 in physical education to 2 standard deviations in medicine. The results for medicine suggest that this program takes highly academically prepared students, so they do not gain much in terms of general knowledge, but there were substantial gains in the specific subject area component.

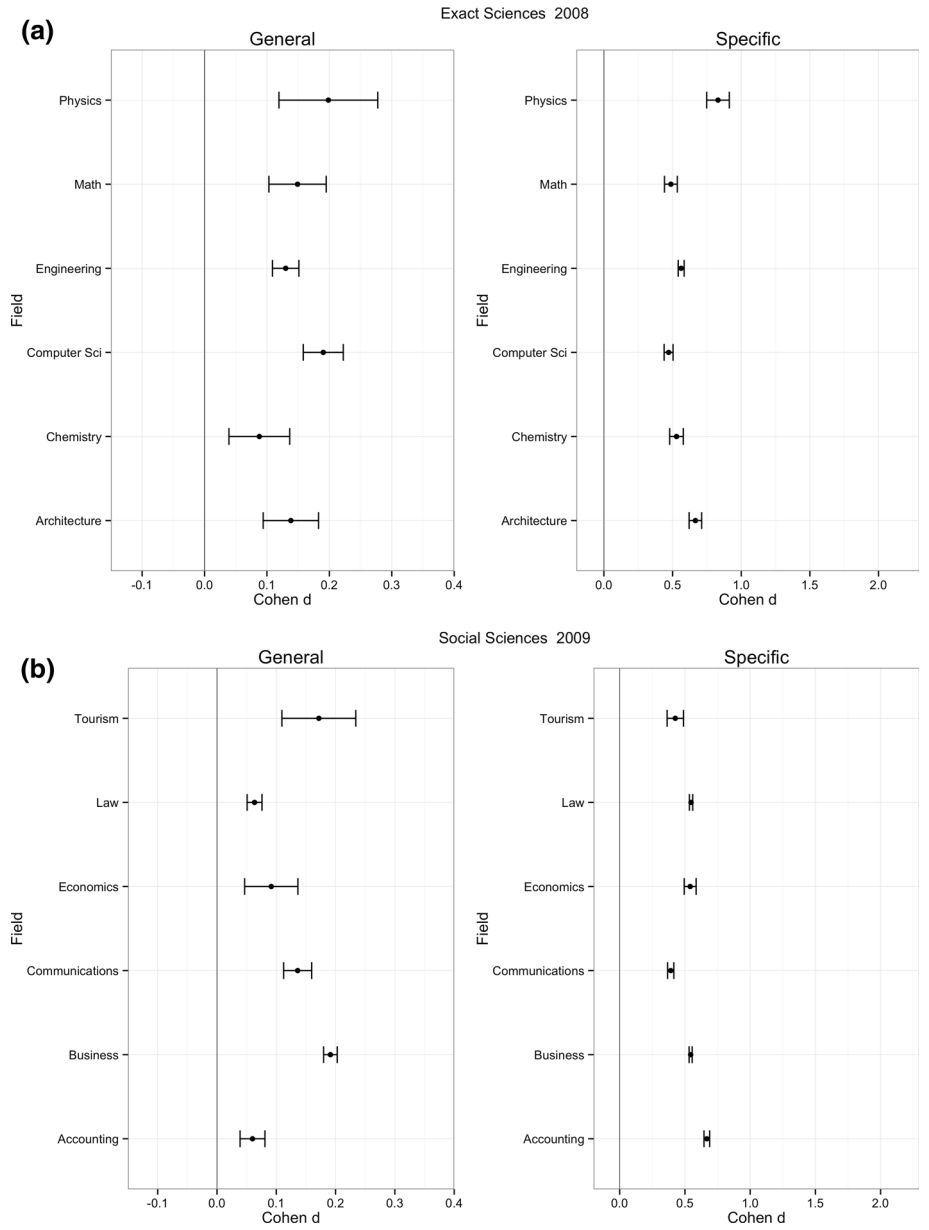

Fig. 1 Gains in average scores in the general and subject area components of ENADE in terms of effect sizes: STEM (**a**), Social Sciences (**b**), and Biological Sciences (**c**)

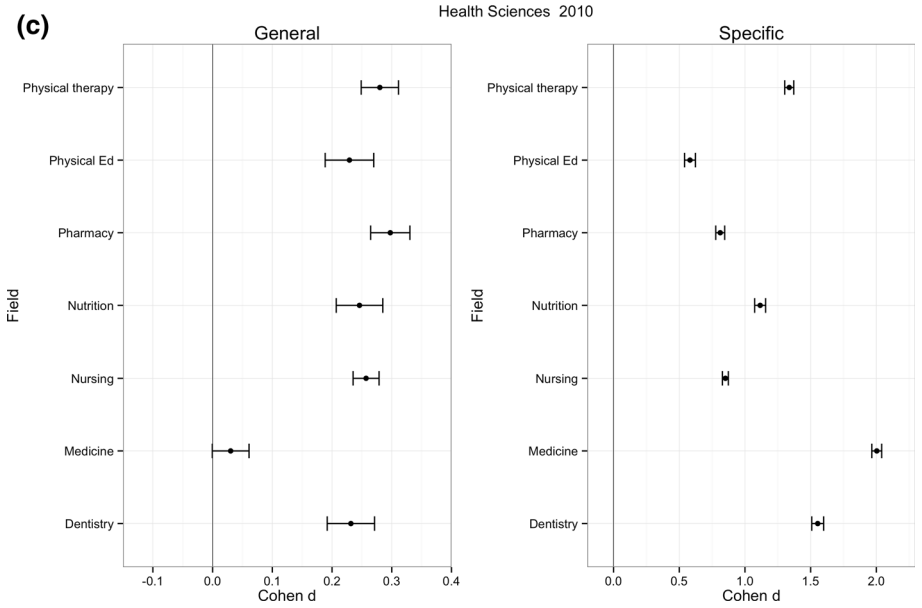

Fig. 1 continued

Gains in SLOs in the general and subject area components for students from the top and bottom income level

The previous results clearly illustrate that there seemed to be larger gains for the students in the *subject area* component compared to the *general knowledge* one. This was not surprising given that the *general* component was not aligned with the curricula of the programs of study. We were also interested in testing whether there was some variation in gains between students of low and high income (Fig. 2a–c). The results were consistent with the one in Fig. 1a–c. It was noteworthy that there were no major differences (with a couple of exceptions in law, pharmacy, and physical therapy) in the gains exhibited by either low- or high-income students.

The previous results suggested that there were no clear patterns in terms of the gains for students from different income levels by major field of study. However, this did not mean that there were no differences in the overall scores between low- and high-income students who took the test. In order to test this, we computed the effect size of the gain for *all* students, as a whole, instead of separating them into major fields of study. The bar indicated the 95 % confidence interval on the effect size, so if there was no intersection in the confidence intervals, there was a significant difference between the effect sizes (with 95 % confidence). The results in Fig. 2d clearly show that for the combined major fields of study the pattern of relatively higher gains in the subject area part, as opposed to the general part of the examination, prevailed. It was also worthy to note the lack of statistically significant differences when dividing the sample between the proportion of students from either low- or high-income backgrounds. Finally, there was substantial variation in the scores of

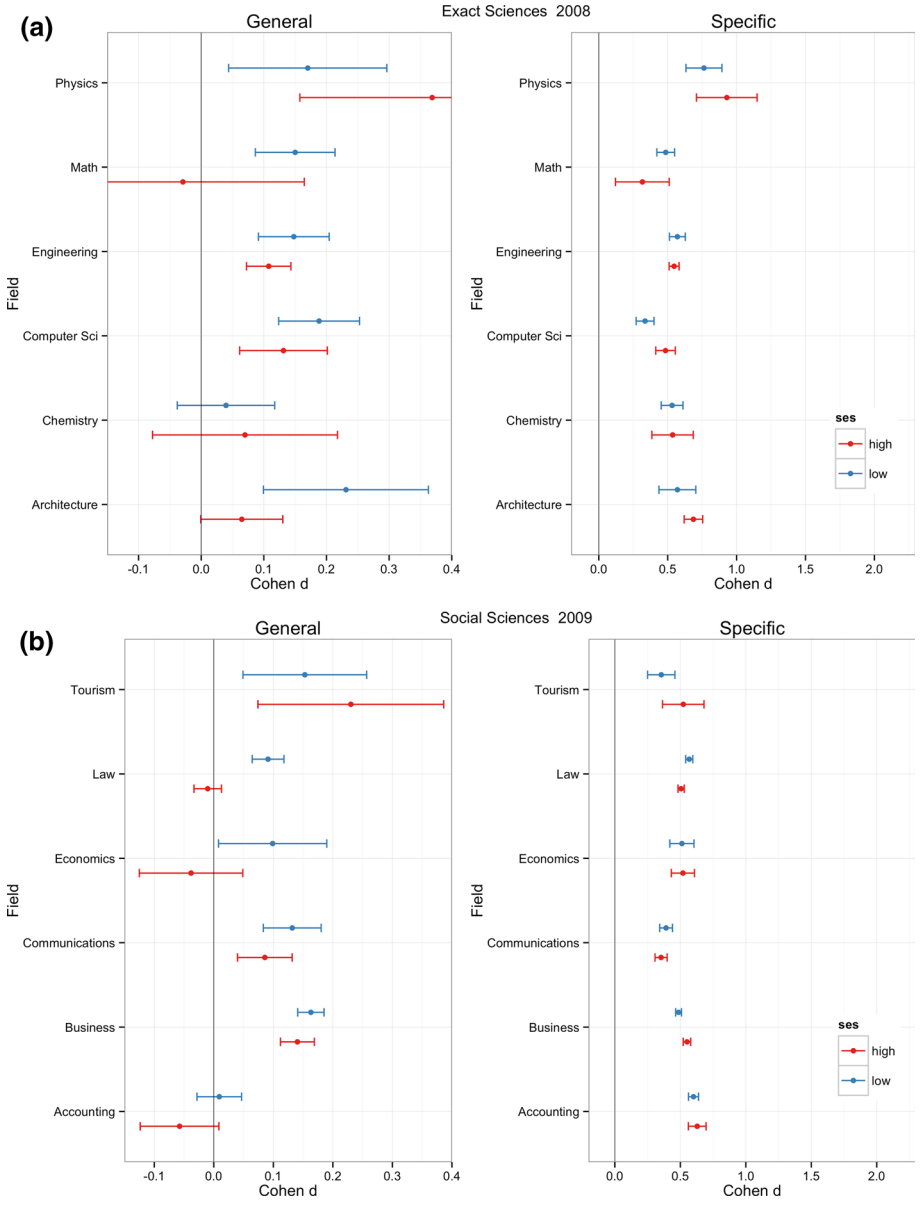

Fig. 2 Gains in average scores in the general and subject area components of ENADE in terms of effect sizes by lowest and highest income levels: (**a**) STEM, (**b**) Social Sciences, (**c**) Biological Sciences, (**d**) all major fields of study

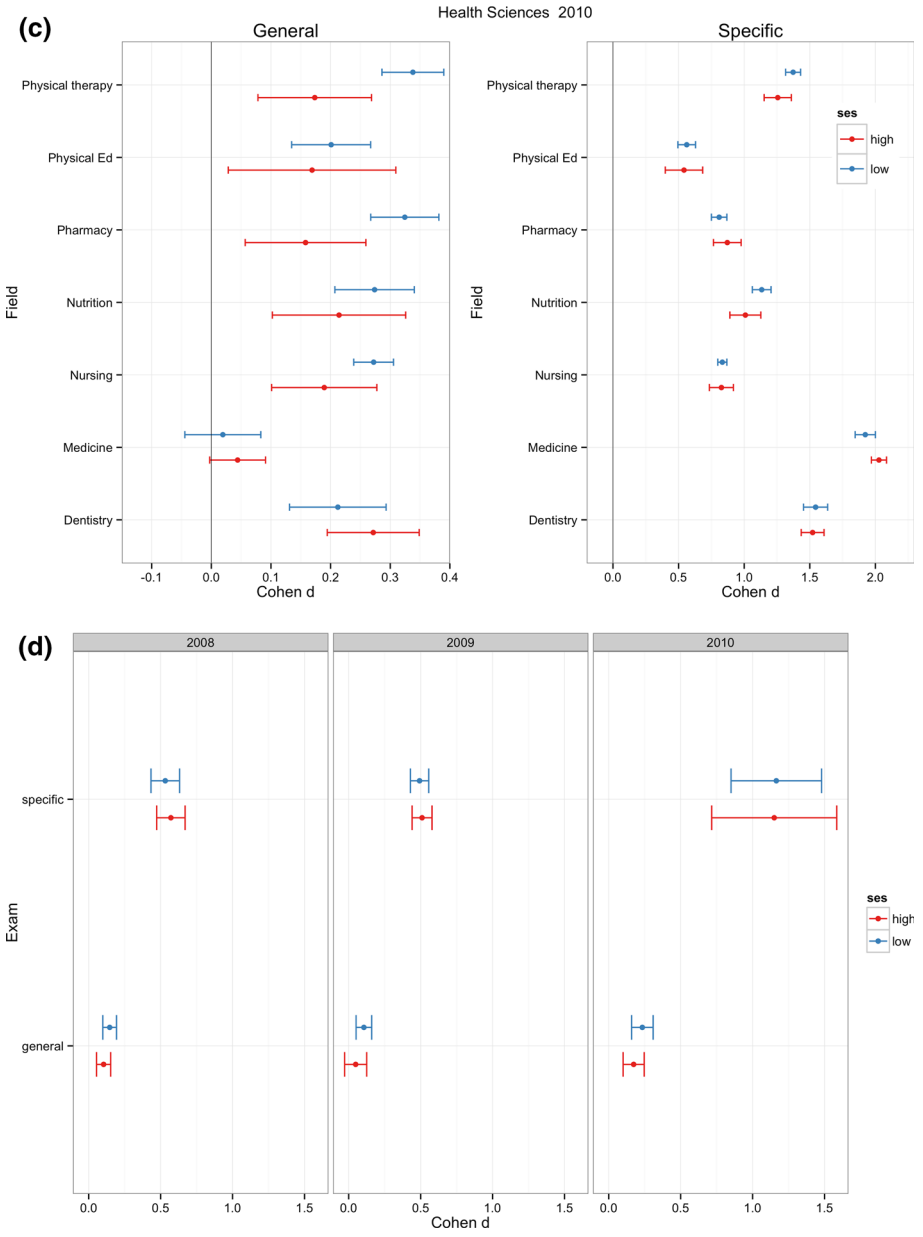

Fig. 2 continued

students attending programs in the Biological Sciences compared to their peers in the STEM and Social Sciences.[20]

Gains in SLOs in the general and subject area components of the test by institutional control

We tested whether the differences in effect sizes varied by institutional control (i.e., public vs. private institutions). The results in Fig. 3a–c show some differences for specific major field of study. For students in STEM programs, it was noteworthy that there were basically no differences in the gains between students from private and public institutions. It was also difficult to make any generalizations in terms of the variability in scores between these two groups of institutions. There was a lot of variation in the scores of the general part for students in physics in the private institutions compared to public ones. The opposite is true for students enrolled in computer science programs; the range in scores was much wider in the public than in the private institutions. One program that stood out was engineering where the gains in the *general component* were much higher for students attending private institutions. The same randomness in the patterns remained for students in the Social Sciences programs (Fig. 3b). There was a wide variation in the *general* component scores and very little in the *subject area* scores. There was also no clear pattern in terms of institutional control as students in tourism programs in public institutions gained a lot compared to their peers in private institutions. However, students in accounting, business, and law enrolled in programs in private institutions gained more compared to the ones in public institutions. Finally, in the programs of the Biological Sciences there were no observed differences in the gains in the general component by institutional control (Fig. 3c). However, it was noteworthy that students enrolled in physical therapy and medicine programs at public institutions gained much more in terms of the *subject area* component, compared to their peers at private institutions.

Figure 3d shows the results of combining the students in all areas. In general, students in private universities achieved larger gains than students in public universities, with the exception of the subject area component of the examination in Biological Sciences.

Conclusions and policy implications

The results of this study provided empirical evidence that students were gaining both general and subject area knowledge in most of the programs of the STEM, Social Science, and Biological Sciences offered by both public and private institutions in Brazil. The results illustrated that there appears to be larger gains in terms of the subject area compared to the general knowledge one. We also found that the majority of the students enrolled in the Biological Sciences fields (with the exception of medicine in the general component) gained more in terms of both general and subject area knowledge than those in STEM and Social Sciences. Interestingly, we found no major differences in gains for students from the

[20] The much higher variability in the subject area part of the examination in 2010 might be related to the substantially higher variation in the estimates for the students enrolled in the different Biological Science programs. This in turn might be related to the way the different tests were created and therefore might not be related to the specific fields or programs. This is a key issue that needs to be taken into account when using these types of methods to compare gains in SLOs in different programs. We thank our reviewer for suggesting this alternative explanation.

highest and lowest income levels when compared by major fields of study. Finally, we could not discern a clear pattern by institutional control by major fields of study.

The findings of this study were in line with the results of Rossefsky-Saavedra and Saavedra (2011) for a single cohort and analogous sample of students (i.e., two different

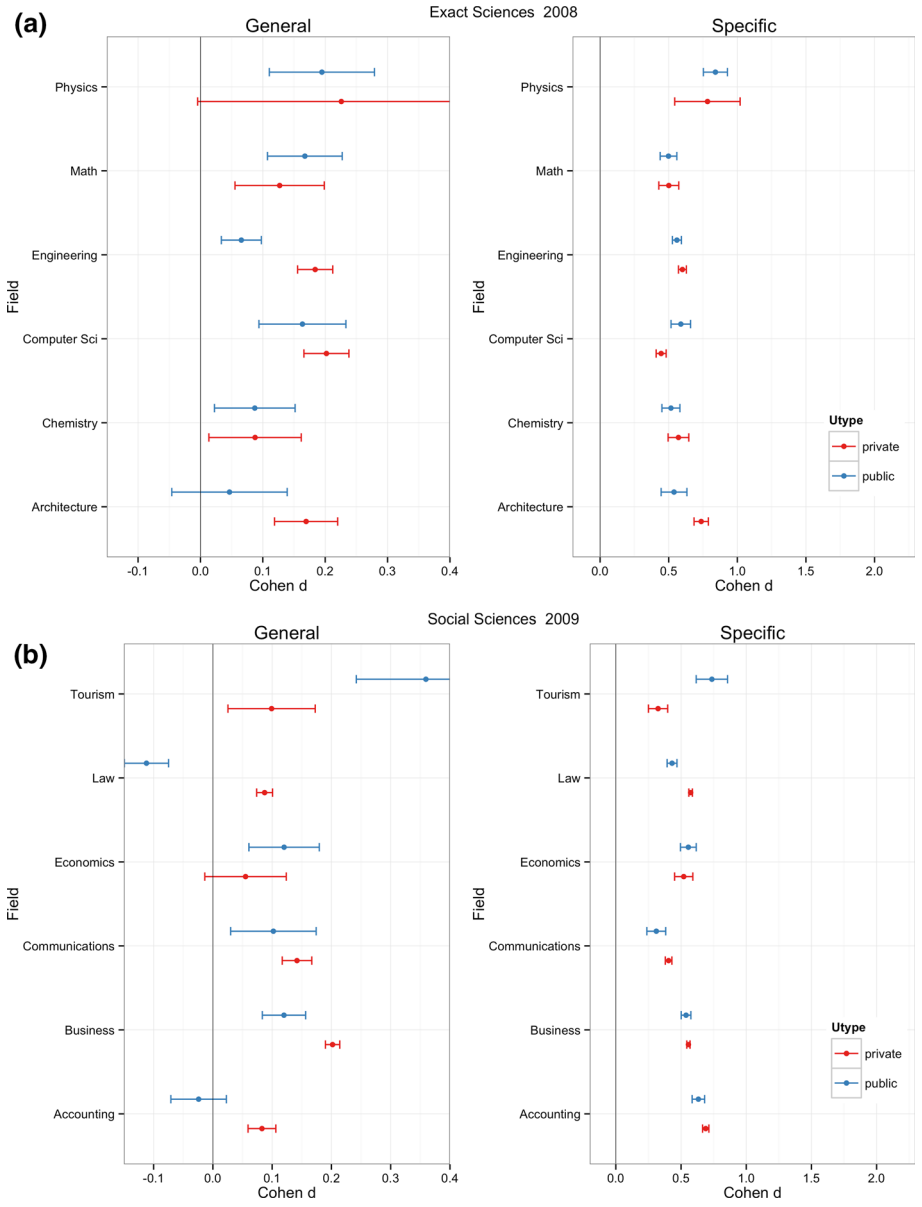

Fig. 3 Gains in average scores in the general and subject area components of ENADE in terms of effect sizes by institutional controls: (**a**) STEM, (**b**) Social Sciences, (**c**) Biological Sciences, (**d**) all major fields of study

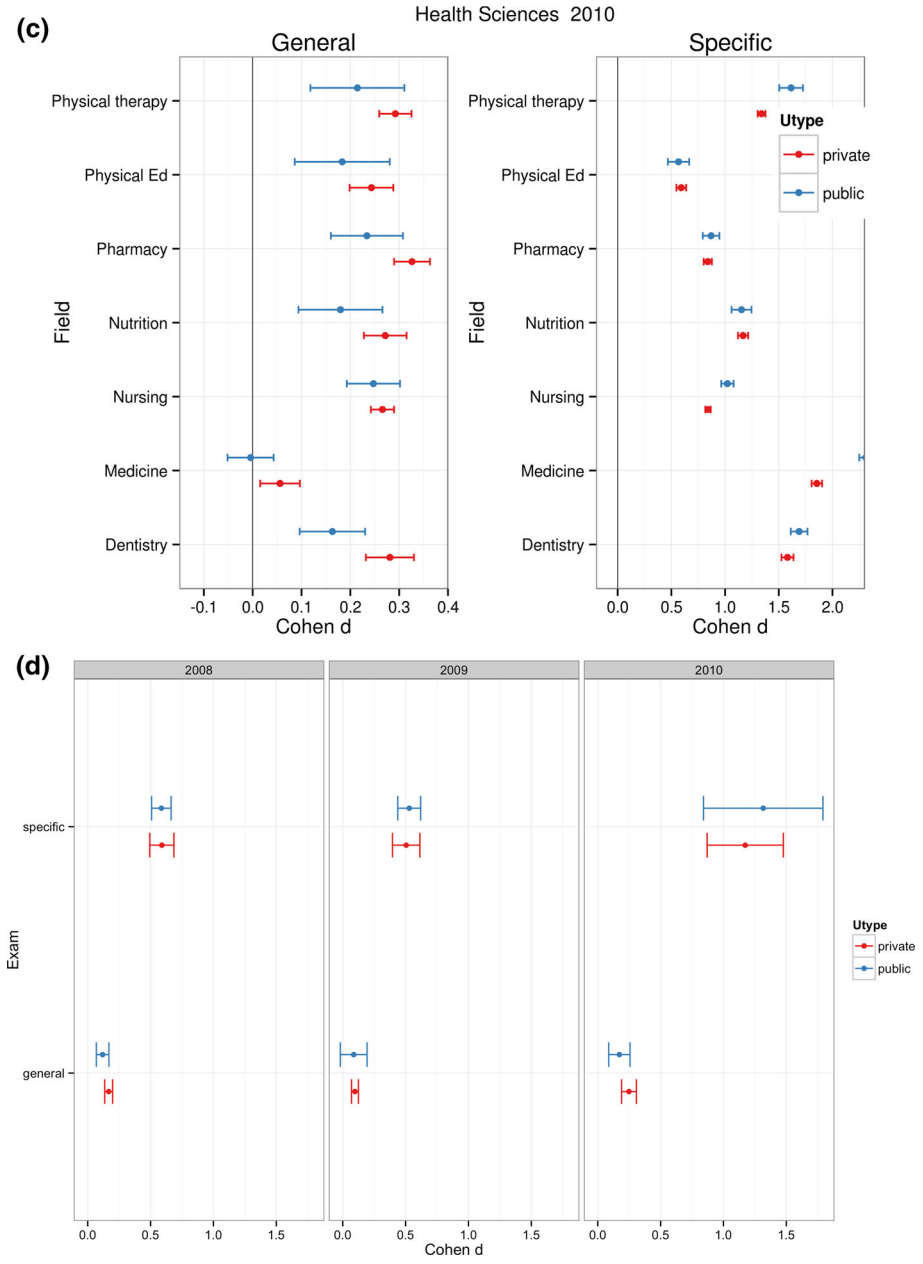

Fig. 3 continued

samples of freshmen and seniors enrolled in the same programs) who participated in a pilot study for the development of a college-exit examination in Colombia. In our case, we found gains for observationally similar students in the two components of the test: the general and subject area knowledge. Even though these studies used different types of tests

that were not measuring the same competencies, it was important to note that students were indeed gaining knowledge and skills. These findings differed from the work of Arum and Roksa (2011). These contradictory findings illustrated the problems of estimating models to measure the gain in SLOs without controlling for the issues of selection of students into institutions and the non-random attrition of seniors (Domingue et al. 2014; Melguizo et al. 2015). Future studies that continue to build in this emerging literature and attempt to produce unbiased measures of gains in SLOs in terms of learning in college, need to use appropriate instruments and methods to address the methodological issues embedded in these types of study. Some recommendations include: (1) choose a college-level test with content that is aligned to the programs of study being evaluated, (2) use a college-level test that ideally has some type of consequences, so the students take it seriously, (3) use appropriate statistical techniques to control for factors associated with college persistence and attainment (i.e., previous academic preparation and non-cognitive factors), and (4) address the issue of non-random attrition of students, especially in the first two years, which is when most of the dropouts take place.

The results of this study have important policy implications for countries interested in developing comprehensive systems to evaluate the quality of higher education institutions. First, the USA could learn from the experiences of Brazil and Colombia and engage in a long-term process of developing a comprehensive evaluation system (Coates 2014). The USA should avoid simply trying to develop a ranking system like the one developed by *U.S. News and World Report*. Second, as countries start to develop appropriate instruments to measure the general and subject area knowledge gained by students, they should work with researchers and testing companies to identify the appropriate instrument to measure the type of competencies that they are interested in measuring. Third, as countries create datasets that can be used to measure the growth in SLOs, researchers, policy makers will have to deal with the methodological problems inherent to these types of studies such as non-random attrition of students. Fourth, governments should also work toward creating a K-20 data system, so they have ample variables to control for previous academic preparation and non-cognitive factors associated with college persistence and attainment. Finally, the information from growth in SLOs should be used in a formative way and comparisons among institutions should be avoided. As documented in the pilot studies of AHELO, it is very important that the information be given back to the institutions as a way for them to continue to work toward improving the students' learning outcomes.

Acknowledgments We would like to thank Roberto Verhine of Universidad Federal de Bahia, Marcelo Knobel and Renato Pedrosa at University of Campinas for helpful comments and suggestions on earlier versions of this paper.

Appendix: Combining effect sizes to produce program-level/engineering effect sizes

The method to combine measures of effect sizes of different fields (e.g., mathematics and computer science) into a general area (e.g., STEM) measure comes from procedures developed in the meta-analyses field.[21] There are two approaches that are generally used to combine effect sizes: fixed and random effects models (Borenstein et al. 2007). The fixed

[21] For the computation of the combined effect size, we used the raw scores of the student in the multiple-measures part of the general examination (NT_OBJ_FG) and the raw score on the essay component of the examination (NT_OBJ_CE).

Table 3 Effect sizes for the general and specific examinations, for all students, for students of high and low income, and for students in private and public universities, for all major field of study

Year	Area	Major field of study	All students		Student income				University type			
					Low income		High income		Public		Private	
			General	Specific	General	Specific	General	Specific	General	Specific	General	Specific
2008	Sciences	Mathematics	0.15	0.49	0.15	0.49	−0.03	0.32	0.17	0.50	0.13	0.50
		Physics	0.20	0.83	0.17	0.76	0.37	0.93	0.19	0.84	0.23	0.78
		Chemistry	0.09	0.53	0.04	0.53	0.07	0.54	0.09	0.57	0.09	0.52
		Architecture	0.14	0.67	0.23	0.57	0.06	0.69	0.17	0.74	0.05	0.54
		Computer science	0.19	0.47	0.13	0.48	0.19	0.34	0.20	0.44	0.16	0.59
		Engineering	0.13	0.56	0.11	0.55	0.15	0.57	0.07	0.56	0.18	0.60
2009	Social Sciences	Business	0.19	0.54	0.16	0.49	0.14	0.55	0.20	0.56	0.12	0.54
		Law	0.06	0.55	0.09	0.57	−0.01	0.51	0.09	0.57	−0.11	0.43
		Communications	0.14	0.39	0.13	0.39	0.09	0.35	0.14	0.40	0.10	0.31
		Economics	0.09	0.54	0.10	0.51	−0.04	0.52	0.06	0.52	0.12	0.56
		Accounting	0.06	0.67	−0.06	0.63	0.01	0.60	0.08	0.69	−0.02	0.63
		Tourism	0.17	0.43	0.23	0.52	0.15	0.35	0.10	0.32	0.36	0.74
2010	Health Sciences	Dentistry	0.23	1.55	0.27	1.52	0.21	1.54	0.16	1.69	0.28	1.58
		Medicine	0.03	2.00	0.04	2.03	0.02	1.92	−0.00	2.31	0.06	1.85
		Pharmacy	0.30	0.81	0.32	0.81	0.16	0.87	0.23	0.87	0.33	0.84
		Nursing	0.26	0.85	0.27	0.83	0.19	0.83	0.27	0.84	0.25	1.02
		Nutrition	0.25	1.12	0.27	1.13	0.21	1.01	0.27	1.17	0.18	1.15
		Physical Ed	0.23	0.58	0.20	0.56	0.17	0.54	0.24	0.59	0.18	0.57
		Physical therapy	0.28	1.34	0.17	1.26	0.34	1.37	0.29	1.34	0.21	1.61

effects model assumes that the effect size of each clinical trial, in our case each field, is an estimate of a true and unique *correct* effect size. Each effect size is different because of sampling error, and thus an *average* of the measures is a good estimate of this combined effect size. These models traditionally use a weighted average because clinical trials that involve more people should weigh more than trials with less people involved. The weight used is the inverse of the variance of the effect size, which is calculated as

$$w_i = {}^{1}\!/_{\sigma_i} \text{ and } \sigma_i^2 = \frac{ne_i + nc_i}{ne_i nc_i} + \frac{d_i^2}{2(ne_i + nc_i)}$$

where *ne* and *nc* are the number of people in the experimental and control groups and the subscript *i* refers to each field's measures. The combined effect size is then calculated as

$$d = \frac{\sum w_i d_i}{\sum w_i}$$

The variance of the combined effect size is calculated as

$$\sigma_i^2 = {}^{1}\!\Big/\!_{\sum w_i}$$

and from this, one can calculate a confidence interval for the resulting effect size.

The random effects model does not assume that all trials, or in our case, all field have the same "true" effect size, but instead assumes that the true effect size for each field is normally distributed around a mean true measure, with a variance, called variance between treatments, denoted by τ (tau). The statistical computations estimate not only the mean true value, but also the between-treatment variance. We refer the reader to Borenstein et al. (2007) for a more complete explanation and the specific formulas. In this study, we chose the random effect model because it is clear from Fig. 1a–c that each field has a different "true" effect sizes, sometimes statistically significantly different. Specifically, we use the DerSimonian–Laird estimation for the between-study variance as implemented in the package *meta* of the software R (Table 3).

References

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

Barnett, R. (1992). *Improving higher education. Total quality care*. The Society for Research in Higher Education and Open University Press, Buckingham.

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 Internaciónal ICFES sobre Investigacion en la Calidad de la Educación,* Bogotá, Colombia.

Borenstein, M., Hedges, L., & Rothstein, H. (2007). *Meta-analysis: Fixed effect vs. random effects*. http://www.meta-analysis.com/downloads/Meta-analysis%20fixed%20effect%20vs%20random%20effects.pdf. Accessed 25 January 2014.

Clark, B. R. (1983). *The higher education system: Academic organization in cross-national perspective*. Berkeley, CA: University of California Press.

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), 1–13.

Coates, H. (2014). *Higher education learning outcomes assessment: International perspectives*. Frankfurt: Peter Lang.

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.
Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28.
INEP. (2009). SINAES: da concepção à regulamentação [SINAES: From conceptual development to legislation], Instituto Nacional de Estudos e Pesquisas Educacionais "Anísio Teixeira" (INEP), Ministério da Educação, Brasil. www.publicacoes.inep.gov.br/detalhes.asp?pub=4389#.
Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20*(8), 1–24.
Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The Collegiate Learning Assessment facts and fantasies. *Evaluation Review, 31*(5), 415–439.
Melguizo, T. (2011). A review of the theories developed to describe the process of college persistence and attainment. In J. C. Smart & M. B. Paulsen (Eds.), *Higher Education: Handbook of Theory and Research* (pp. 395–424). The Netherlands: Springer.
Melguizo, T., Zamarro, G., Velazco, T., & Sanchez, F. (2015). *How can we accurately measure whether students are gaining valuable learning as well as other relevant outcomes in Higher Education?* Rossier School of Education, University of Southern California.
National Academy of Academic Leadership. (2014). *Assessment and evaluation in higher education: Some concepts and principles*. http://www.thenationalacademy.org/sitemap.html.
Nusche, D. (2008). *Assessment of learning outcomes in higher education. A comparative review of selected practices,* OECD education working papers no 15, OECD, Paris.
OECD. (2008). *Higher education to 2030,* Vol. 1, Demography. Centre for educational research and innovation. Organization for Economic Cooperation and Development.
OECD. (2013). Assessment of higher education learning outcomes (AHELO): Feasibility study report. Volume 1: Design and implementation. Retrieved on 8 April 2014 from http://www.oecd.org/education/skills-beyond-school/testingstudentanduniversityperformancegloballyoecdsahelo.htm.
Pedrosa, R. L., Amaral, E., & Knobel, M. (2013). Assessing higher education learning outcomes in Brazil. *Higher Education Management and Policy,*. doi:10.1787/hemp-24-5k3w5pdwk6br.
Possin, K. (2013). A serious flaw in the Collegiate Learning Assessment [CLA] test. *Informal Logic, 33*(3), 390–405.
Primi, R., Carvalho, L. F., Miguel, F. K., & Silva, M. C. R. (2010). Análise do funcionamento diferencial dos itens do Exame Nacional do Estudante (ENADE) de Psicologia de 2006. [Analysis of the differential item functioning of the 2006 Psychology ENADE exam]. *Psico-USF 15*(3), 379–393. http://www.scielo.br/scielo.php?pid=S1413-82712010000300011&script=sci_arttext.
Primi, R., Hutz, C. S., & Silva, M. C. R. (2011). A prova do ENADE de Psicologia 2006: concepção, construção e análise psicométrica da prova. [The 2006 Psychology ENADE exam: conception, construction, and psychometric evaluation of the exam]. *Avaliação Psicológica 10*(3), 271–294. http://www.labape.com.br/labape/artigos/A%20PROVA%20DO%20ENADE%20DE%20PSICOLOGIA%202006.pdf.
Rossefsky-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.
Saavedra, J. E. (2009). *The learning and early labor market returns to college quality: A regression discontinuity analysis*. Cambridge, MA: Harvard University.
Silva Filho, R. L., et al. (2007). A evasão no ensino superior brasileiro. [Evasion in Brazilian higher education]. *Cadernos de Pesquisa, 37*(132), 641–659.
Steedle, J. T. (2012). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education, 37*(6), 637–652.
Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics, 25*(1), 1.
Verhine, R. & Dantas, L. M. V. (2005). Assessment of higher education in Brazil: from the provão to Enade". Document prepared for the World Bank, responsible party: Alberto Rodriguez.
Verhine, R., Dantas, L. M. V., & Soares, J. F. (2006). Do Provão ao ENADE: uma análise comparativa dos exames nacionais utilizados no Ensino Superior Brasileiro [From "Provão" to ENADE: A comparative analysis of national exams used in Brazilian higher education]. *Ensaio: Aval. Pol. Públ. Educ.*, *14*(52), 291–310. See earlier English version.
INEP. (nd). Resultado do indicador de diferença entre os desempenhos observado e esperado—IDD [Results from the indicator of the difference between the observed and expected outcomes], Instituto Nacional de Estudos e Pesquisas Educacionais "Anísio Teixeira" (INEP), Ministério da Educação, Brasil.
Zemsky, R., Wegner, G. R., & Massy, W. P. (2005). *Remaking the American University: Market-smart and Mission-centered*. Piscataway, NJ: Rutgers University Press.