Current Issues in Education
Mary Lou Fulton Teachers College • Arizona State University
PO Box 37100, Phoenix, AZ 85069, USA

Volume 18, Number 1 March 22, 2015 ISSN 1099-839X

Assessment under Resource Constraints

Steve Lovett and Mary G. Curtis
The University of Texas at Brownsville

Assessment and the measurement of learning are receiving increasing emphasis in American higher education. This is a case study that demonstrates a simple, inexpensive method of measuring freshman to senior "gains" or learning using a cross-sectional methodology. Seniors and freshmen within a four-year business program were both given the same multi-part test. Not surprisingly, the seniors' average score on all parts was higher than that of the freshmen. However, the seniors were older than the freshmen, indicating a possible maturity effect, and had higher average scores on their entrance examinations, indicating a possible selection effect. We used regression techniques to estimate these effects, and subtracted the estimate from the seniors' gain to estimate a net gain. Our method is applicable to any learning outcome that can be quantified, and we believe that it is both effective and within the means of nearly all U.S. institutions of higher education.

Keywords: Program assessment, measuring learning, cross-sectional design, Hispanic-serving institution, critical thinking skills

Institutions of higher education in the United States are becoming more and more aware of the need to measure and manage student learning. Zumeta referred to an "outcomes revolution" (2011, p. 137) in American higher education, which began sometime in the 1980s and in which educational programs are evaluated more on the basis of results, such as measures of student learning, job placement rates, or graduates' career success, and less on the basis of inputs, such as faculty credentials, advanced scientific laboratories, or library facilities. This shift is reflected in the priorities of accrediting agencies. Banta, Jones, and Black noted that "Most professional accrediting organizations expect faculty within accredited academic programs to demonstrate accountability regarding student performance on a continuous basis" (2009, p. 22). Still, Rodgers et al. (2013) suggested that seeking real improvements in student learning is an even more powerful rationale for program assessment than simply responding to the pressures of accrediting agencies or any other external stakeholders.

It's useful to distinguish institutional goals and outcomes from program goals and outcomes. Goals at both levels should be mission driven (Council for Higher Education Accreditation, 2014). Institutional mission statements tend to emphasize broad measures of success and contributions to communities. For example, the institution at which this study was conducted is located on the U.S./Mexican border, and its mission is focused on the integration of cultures and on developing knowledgeable citizens and leaders. However, this paper is focused on program goals and outcomes. College, school or program mission statements are generally more specific than those of institutions. For example, this study was done within the business school of the above mentioned institution, and the school mission emphasizes teaching critical thinking, quantitative skills, and business content knowledge, all of which we attempt to measure as part of the program assessment process.

Unfortunately, program assessment is often met with resistance from faculty. Several authors, including Bresciani (2006), Suskie (2009) and Walvoord (2010),

have suggested that resource constraints are a key reason for this resistance – assessment requires resources, including faculty time and labor, and administration is often either reluctant or unable to provide these resources. Suskie wrote ". . . if assessment is to be pervasive and sustained, it must be cost-effective, yielding benefits that justify the time and expense put into it" (2009, p. 90). Therefore, the development of assessment methods that make efficient use of available resources should be a priority within the industry of higher education.

Program assessment is a cycle that begins with setting goals for student learning, continues with teaching or otherwise providing learning opportunities, then with the measurement of learning, and finally with some change in curriculum or pedagogy intended to improve learning (Suskie, 2009). This paper focuses on the third step of the cycle: the measurement of learning. It is a case study, the purpose of which is to demonstrate a simple yet effective method of measuring learning that is within the means of nearly all institutions, including those with limited resources to devote to assessment. Furthermore, this method is versatile in that it is applicable to any learning outcome that can be quantified; in other words, so long as a number can be assigned to a student's result.

Longitudinal versus Cross-Sectional Designs

The two basic designs for measuring learning are longitudinal and cross-sectional. A longitudinal design means that the same student is evaluated at different times – for example, as an entering freshman and then again as a graduating senior. The difference between the student's scores in whatever measure is used is taken as the student's "gain" while in the program. A cross-sectional design means that different students are evaluated at the same time. For example, in one semester a group of entering freshmen and another group of graduating seniors could be evaluated, and the difference between their scores could be taken as the freshman to senior gain.

Seifert et al. (2010) favored longitudinal designs over cross-sectional designs, writing ". . . longitudinal pretest-posttest designs yield the most internally valid results and the most accurate estimates of college impact. There is no substitute for the 'gold standard' that longitudinal pretest-posttest studies furnish in accurately assessing how students learn and change" (p. 14). These authors did acknowledge several difficulties with longitudinal designs, especially the increased cost of data collection and difficulties resulting from participant dropout, but they nonetheless recommended longitudinal designs because they yield more accurate results than cross-sectional designs.

However, at many institutions, especially those with low persistence, the problems associated with longitudinal studies are magnified to the point that they become impractical. Such studies require that students be tracked through their programs. This is a manageable task when many students complete their programs as expected, as often tends to be the case at wealthier and more selective institutions. Still, tracking students becomes especially difficult in situations of low persistence, or those in which relatively few students actually graduate from the institution in which they begin their education, many of those who do attend irregularly for long periods of time, and many other graduates are transfers from other institutions. Unfortunately, these situations may in fact be the norm. Pascarella and Terenzini wrote that "Based on evidence of nationally representative samples, it would appear that since the late 1980s, 50 percent or more of the students who initially enrolled at a four-year college eventually attended two or more undergraduate institutions" (2005, p. 146).

Furthermore, the persistence problem tends to be more severe at less selective institutions, and these are often those with fewer financial resources. Poorer institutions have fewer resources from which to draw and rely more heavily on student enrollment to fund themselves. For example, the institution at which this study was conducted is located in one of the poorest areas in the nation (U.S. Department of Commerce, 2012) and does not have access to any significant endowments or other financial resources, so administrators must request funding from the state legislature biannually based on enrollment. At least partly because of this, the institution followed an open enrollment policy at the time of the study.

ACT (2010) reported that the first to second year retention rates for four-year public institutions offering bachelor's degrees averaged 91.5% for the "highly selective" institutions, but only 57.5% for open enrollment institutions. The six year graduation rate for the highly selective institutions was 76.0%, but only 26.4% for the open enrollment institutions. And the persistence problem is often especially severe in Hispanic Serving Institutions, such as the one in which this study took place. The U.S. Census Bureau (2010), the U.S. Department of Education (2011), and the U.S. Department of Labor (2011) have all noted that persistence is a particular problem in minority serving institutions. We recognize that the issue of persistence may be viewed in various ways, and that some institutions or educators may in fact view a lack of persistence positively. They may prefer that students compete for degrees, and even encourage low performers to leave their programs. None-the-less, our position is that persistence is to be encouraged – we want students to get degrees because we believe that doing so will contribute to the quality of their lives. As many of the students enrolled at this institution are first-generation college students, persistence to the completion of a four-year degree is a significant challenge. They have few role models within their own families or even within the community.

In any case, a solution to the difficulties that low persistence imposes on longitudinal designs may be the

use of cross-sectional designs, in which different students are evaluated at the same time. Of course, the fundamental problem with cross-sectional studies, and the reason that they are less precise than longitudinal studies, is that different individuals are being evaluated, and they may differ on many dimensions, including their initial preparation or abilities. However, Lovett and Johnson (2012) demonstrated techniques whereby defensible adjustments could be made for any identifiable differences. Furthermore, because they require fewer resources, cross-sectional designs are more likely to be realized in less wealthy institutions. The measurement of learning at these institutions is certainly as critical as it is at more elite institutions, and in many cases these institutions do not require the precision needed for high-level academic research, but simply need to obtain reasonable estimates of student learning on a continuous basis. This paper therefore continues with a demonstration of how estimates of student learning can be obtained through a cross-sectional design, using a simple, pragmatic method that is within the means of nearly all institutions.

The Setting, Participants and Procedures

This study was conducted in the business program of a small university located on the U.S./Mexican border. The institution is a designated Hispanic Serving Institution (U.S. Department of Education, 2011). University records showed that the student body at the time of the study was 93% Hispanic. Many students were raising families and/or working, and most were first generation college goers.

The struggle to increase the educational levels of Hispanics in the U.S. has been well documented by authors such as Amaro-Jiménez and Hungerford-Kresser (2013) and Oseguera, Locks, and Vega (2009). Meinert reported that ". . . only 64.3 percent of Hispanics age 25 and older have at least a high school diploma or the equivalent, and just 14.1 percent have a bachelor's degree . . ." (2013, p. 29). She also stated that "Over the next four decades, 37.6 million Hispanic workers are projected to join the U.S. labor force, which will account for about 80 percent of the total growth in the workforce" (Meinert, 2013, p. 30), and made a rather convincing case that improving the educational level of Hispanics is one of the greatest challenges facing the U.S. today.

First, we obtained approval for the study from our Institutional Review Board. The participants were either undergraduate seniors enrolled in the business program's capstone course or freshmen enrolled in an introductory business course. The two groups were compared using an in-classroom test that counted toward a portion of their course grades – it was an embedded assessment tool. Suskie (2009) favored embedded assessments over "add-on" assessments, or those that are ungraded and outside of course requirements, noting that it is difficult to convince students to take add-on assessments seriously. The test was written by the teacher of the capstone course, and so was an internally generated rather than a standardized test. In general, internally generated tests are more useful for measuring learning because they have greater content validity; they are written by the instructors from content covered in the program, and so measure more specifically what students are expected to learn in the program. Choosing subjects or areas for testing is a very significant issue in any attempt to measure learning within a program. The subjects covered in the test described below were chosen based on the school's mission statement, the required classes in the school's degree programs, and the best judgment of the teacher of the capstone course, who had been teaching within the school for more than a decade.

The test included 70 multiple choice questions. Multiple choice tests are often criticized because they provoke little thought, but they none-the-less remain popular because they can provide information about a broad range of material in a relatively short test (Suskie, 2009). The first 54 multiple choice questions were content questions covering the seven business topics most emphasized in the program. These were Accounting, Economics, Finance, Marketing Management, Marketing Promotions, Operations Management and Organizational Behavior. These are referred to below simply as topics 1 through 7 (not in the same order as above). The test also included 16 other multiple choice questions involving mathematical calculations. The first eight of these are referred to below as "rote math". These were rather simple questions, examples of which had been demonstrated in class by each teacher shortly before the test. The second eight questions are referred to as "math with thought." These were more complex questions that had not been demonstrated but which built on the rote math questions. There was also one essay question which the authors used to evaluate critical thinking and which is discussed more fully below.

There were 49 seniors, constituting the entire enrollment of two sections of the program's capstone course. These seniors took their test in one sitting. There were 177 freshmen enrolled in four sections of the introductory course. Enrollment in these sections ranged from 41 to 47 students, and two different teachers taught two sections each. Both teachers agreed to help with this study, and, in order to minimize the demands on each, we split the task of measurement at the freshman level between them. This meant that not all freshmen participated in all parts of the study, but we consider this to be a realistic assessment situation. Good assessment depends on an efficient use of resources, and accurate estimates of student learning do not require that all students participate in all assessment activities. Suskie wrote that ". . . a representative sample can yield information that is almost as accurate as information from everyone" (2009, p. 46), and noted that sampling is a way

to minimize the costs of assessment. We address the issue of the representativeness of the freshman sections below, at the end of the initial analysis section.

The first freshman teacher had two sections with a total of 89 students. He/she included all of the content questions on in-classroom tests, but the students did not answer all of the questions in one sitting. Because these questions were designed for seniors they were rather difficult for the freshmen, so the teacher included them on tests throughout the semester as the topics came up, and mixed them with easier questions that play no role in this study. Students are usually accustomed to tests that include some easy and some harder questions, so this resulted in rather typical freshman level tests. The second freshman teacher had two sections with a total of 88 students and included the math questions on tests throughout the semester in a similar fashion, and also included the critical thinking essay on a test about halfway through the semester.

It is worth noting here that the baseline used for measuring gains in this study was the freshmen's scores *during* an introductory course, after they had received some instruction. Of course, the ideal baseline would be the freshmen's "pre-entry" scores, or their score *before* entering the program. However, obtaining pre-entry scores using an embedded assessment instrument, or one that counts as a part of a course grade, may simply be impractical. Doing so would mean giving the freshmen a graded test that they had no opportunity to study for on or before their first day of class. The freshmen would almost certainly perceive this to be unfair, and many teachers would agree with them.

Measuring Underlying Constructs

Of all of the outcome variables used in this study, critical thinking (CT) was the most difficult to measure. Halpern (2003) defined CT as ". . . thinking that is purposeful, reasoned, and goal directed" (p. 6), and there is little doubt that it is a priority for teachers in higher education. Buskist and Irons (2008) wrote "If there is one thing that all college and university teachers want their students to learn, it is to think critically" (p. 49), and Halpern (2001) wrote that ". . . there is virtually no disagreement over the need to help college students improve how they think" (p. 270).

Bailin et al. (1999) pointed out one aspect of CT that is especially relevant to this study; CT requires some background knowledge of the issue at hand. They wrote that ". . . the depth of knowledge, understanding and experience persons have in a particular area of study or practice is a significant determinant of the degree to which they are capable of thinking critically in that area" (Bailin et al., p. 290). For example, a person who knows little or nothing about chemistry is at a severe disadvantage if asked to apply CT to determine whether the results of a chemistry experiment are valid. Likewise, a person who knows little about business is at a disadvantage if asked to apply CT to a business situation. This by itself was good reason to hope that the seniors in our study would demonstrate better CT skills than the freshmen.

Both the seniors and freshmen were asked to write a short essay from the same prompt that presented the students with a business situation in which a low-end U.S. shoe manufacturer had to decide whether to begin manufacturing "knock-offs" of an expensive Italian shoe. This situation had many points to consider. For example, while the Italian shoes might sell well in some markets, they might not sell well in the markets of the low-end manufacturer. There was also the response of the Italian company to consider – would they take legal action? – and there were ethical issues to take into account as well. This prompt was one that Suskie (2009) would consider an "extended response" rather than a "restricted response" prompt because it gave the students considerable latitude in deciding how to respond, which was appropriate for an essay intended to evaluate critical thinking.

The essays used for scoring were handwritten. The course teachers first made clean photocopies of the essays. We then shuffled these so that we wouldn't know whether the student was a senior or a freshman until after we had assigned a final score, scored each essay independently, and finally met to compare and discuss each student's essay in order to come to a consensus on the final score. We used a descriptive rubric for scoring. This means one that includes brief descriptions of the work that would result in each possible rating. Suskie (2009) noted that using rubrics as scoring guides makes scoring easier and more consistent, and that descriptive rubrics are especially useful because they explicitly document standards of performance. Using a 3 point scale (3 = Exemplary, 2 = Competent, 1 = Developing), we scored each essay on four CT competencies. We evaluated how well each student: 1) restated the problem described in the prompt, 2) analyzed the issues in the prompt, 3) used synthesis, or supported their response with external information, and 4) came to a logical conclusion or evaluation based on the response. We used a pilot group of essays not included in the project results to clarify and refine the rubric. The final CT score was the simple or unweighted sum of the four competency scores.

All of the outcome measures used in this study were therefore *direct* evidence of student learning, which Suskie describes as ". . . tangible, visible, self-explanatory, and compelling evidence of exactly what students have and have not learned" (2009, p. 20). Also, all of the measures used were *quantitative*, because they resulted in meaningful numbers that could be analyzed statistically. However, the multiple choice questions (content and math) were *objective* measures, while the critical thinking essay was a *subjective* measure, or one that required professional judgment to score. Suskie (2009) noted that subjective measures are becoming

increasingly popular because they evaluate important skills that cannot be evaluated through objective measures, including creativity and problem solving skills.

However, as mentioned above, our cross-sectional methodology required that possible differences between the seniors and freshmen be taken into account. One difference that is likely to be important is a selection effect – students who are less prepared for college are more likely to drop out, while those who are well prepared are likely to persist. This selection effect is an integral part of Tinto's (1975, 1993) model of student departure. It tends to raise the average performance of those remaining in a program, thereby creating a false appearance of learning within the program. We therefore needed to estimate the students' readiness upon entering the program. We obtained the information to do this from the university's admissions center. All of the students had taken commercially available reading, writing and mathematics tests upon entering the university. We standardized the scores from these tests, and below refer to the average of the three standardized scores as a "readiness composite" or simply "readiness."

Initial Analysis

Sample means and the results of t-tests for the equivalence of means are shown in Table 1. We refer to the first seven variables as demographic variables, number eight (the readiness composite) as a selection variable, and numbers nine through twelve as performance variables.

Table 1

Descriptive Statistics – Seniors versus Freshmen

	Seniors (n=49)	Freshm. (n=177)	Overall St. Dev.	Sig. of Diff.
1. Avg.. age	27.51	20.84	5.635	.000
2. Male (%)	53.0	55.0	0.499	.775
3. Hispanic	91.3%	94.6%	0.250	.474
4. Speak mostly Spanish at home	34.8%	38.7%	0.484	.711
5. Mother college graduate	26.1%	33.3%	0.459	.494
6. Father college graduate	32.6%	39.3%	0.483	.511
7. Avg. family income	$39,350	$34,930	18,351	.200
8. Readiness composite	0.2576	-0.0563	0.673	.004
9. Content MC (avg. correct)	85.8%	66.3%	0.164	.000
10. Math rote (avg. correct)	90.0%	75.7%	0.190	.000
11. Math with thought (avg. correct)	74.0%	53.4%	0.243	.000
12. Critical thinking (out of 12)	7.67	6.98	1.713	.044

We took the demographic variables from either university records or a student survey done in class. These show that the seniors and freshmen were similar in several ways. In both groups just over half were men, and most were Hispanic. Just over a third reported speaking mostly Spanish at home. Roughly one-third reported that their parents were college graduates – in other words, about two-thirds of both groups were first generation college students. Their average family income was less than $40,000 per year.

But there were differences as well. It was not surprising that the seniors were older than the freshmen, but this was important because it indicated a likely maturity effect. Pascarella and Terenzini wrote: "It is one thing to conclude that increases in subject matter knowledge and academic skills occur *during college*. It is quite another to conclude that these increases occur *because of college*" [emphasis in original] (2005, p. 70). Certainly, college is not the only avenue through which students learn as they go through life – they learn new vocabulary by watching television, for example, and they may learn business skills through work experience. Therefore, in order to calculate a net gain, an adjustment for age may be necessary.

Table 1 also shows that the seniors' readiness composite was higher, indicating a possible selection effect, or one in which poorly prepared students dropped out at a higher rate than well prepared students (Tinto, 1975, 1993). Again, an adjustment for such an effect may be necessary in order to calculate a net gain. Finally, the seniors' scores were higher on all performance measures – the content multiple choice (topics 1-7 combined), rote math, math with thought, and the critical thinking essay. Of course, this was an initially encouraging result, although it was before adjustments for differences between the groups.

Table 2 shows correlations for the combined data set including both seniors and freshmen. Several correlations were noteworthy. Age correlated negatively with speaking Spanish at home and with parents' education levels – this probably reflects the increasing use of English and increasing education levels within the area. The men reported higher family incomes, and the Hispanics lower family incomes. It was interesting that those who spoke Spanish at home tended to report higher parents' education levels – perhaps the Spanish speakers were the more recent arrivals in the area and perhaps education was related to mobility.

However, our goal at this point was to identify variables for which adjustments should be made when comparing the seniors' and freshmen's performance measures. An adjustment should be made for any variable for which there is a significant difference between the seniors and freshmen, and which correlates significantly with one or more of the performance measures (Lovett & Johnson, 2012). Age was an obvious choice – the seniors were significantly older and age showed a significant positive correlation with three of the four performance measures. However, no adjustments were necessary for any of the other six demographic variables because, even

Table 2
Correlations

	1	2	3	4	5	6	7	8	9	10	11
1. Age											
2. Male	-.04										
3. Hispanic	-.11	-.16									
4. Spanish/home	-.20*	-.12	.20*								
5. Mother college	-.21*	.09	-.04	.21*							
6. Father college	-.25**	.06	-.07	.36**	.60**						
7. Family income	-.09	.21*	-.25**	-.06	.38**	.42**					
8. Readiness	.08	-.13	-.10	-.27**	-.03	-.15	.09				
9. Content MC	.36**	.18	-.18	-.39**	-.18	-.22	-.21	.42**			
10. Math rote	.27**	-.06	-.08	-.10	.02	-.13	.08	.36**	.44**		
11. Math/thought	.32**	.15	-.25**	-.27**	-.01	-.15	.10	.36**	.42**	.54**	
12. Critical thinking	.17	.11	-.24*	-.16	.16	.17	.19	.28**	.24	.21*	.28**

**= correlation is significant at the 0.01 level (2-tailed).
*= correlation is significant at the 0.05 level (2-tailed).

Table 3

Results: Calculate Net Gains in Ten Areas

	Seniors	Freshmen	Difference	Adjustment	Net gain	Overall std. dev.	Gain in std. dev.
(% correct)							
1. Topic 1	84.69	66.46	18.23	9.17	9.06	24.69	.367
2. Topic 2	87.24	63.72	23.52	9.71	13.81	20.59	.671
3. Topic 3	87.50	67.44	20.06	8.64	11.42	19.56	.584
4. Topic 4	90.09	67.47	22.62	10.66	11.96	22.13	.540
5. Topic 5	84.44	77.73	6.71	3.75	2.96	17.31	.171
6. Topic 6	87.24	65.66	21.58	10.44	11.14	22.90	.486
7. Topic 7	79.85	52.87	26.98	9.83	17.15	26.85	.639
8. Math – Rote	90.05	75.74	14.31	8.63	5.68	19.01	.299
9. Math with Thought	73.98	53.42	20.56	12.08	8.48	24.28	.349
10. Critical Thinking	7.67	6.98	0.69	0.53	0.16	1.713	.093
(Scale = 0 – 12)							

though "Hispanic" and "Spanish at home" correlated negatively with some performance measures, there were no significant differences between the seniors and freshmen on these variables. But the selection variable – the readiness composite – showed significant correlations with all four performance variables, and seniors had higher readiness scores.

We therefore made adjustments for only two variables: age, which represented a maturity effect, and the readiness composite, which represented a selection effect. None-the-less, that the exercise of comparing seniors and freshmen using all available information was a useful one. In this case we found no significant demographic differences between the two groups apart from age, but in other cases we might.

We also compared the freshman groups. Recall that, while each senior participated in the complete study, each freshmen participated in only a part of the study – either content, or math and essay – and the part that they participated in was not the result of a random assignment, rather, an entire section was assigned to the same part. We therefore checked for differences between the groups on age and readiness. The 89 "content" freshmen averaged 21.2 years old, while the 88 "math and essay" freshmen averaged 20.5 years old, and a simple t-test showed that the difference was not significant at traditional levels (p = .30). The average readiness score of the "content" freshmen was -.0236, while that of the "math and essay" freshmen was -.0893, and again a t-test showed no significant difference (p = .53). Therefore, differences between the sections were not a concern. But again, the exercise of comparing the sections was useful because in other cases there might be differences. For example, a night section might have older students, or an honors section might have better prepared students.

Results

The results are shown in Table 3. The first column shows the average senior score on each topic, the second the average freshman score, and the third column the "gain," or the difference between the two scores.

The fourth column shows the adjustment for age and readiness. An example of the procedure for calculating this adjustment is shown in Table 4. In step A, we ran a regression using the topic 1 score as the dependent variable and age and readiness as independent variables, and this regression was done for the entire data set, including both seniors and freshmen. Since the beta values for both of these variables were significant, we know that in general older students and better prepared students tended to score higher on topic 1. Furthermore, the beta values are unbiased estimators of the effect of the age or readiness on the topic 1 score, and this allows us to use the same statistical technique for making adjustments as was used by Lovett and Johnson (2012).

Table 4

Results: Calculate Maturity-Selection Adjustments

Step A - regression on topic 1 score.

Variable	Beta	Std. err.	Sig.
Constant.	.514	.078	.000
Age	.009	.003	.006
Readiness	.101	.034	.004

Model r-square .126

Step B - calculate predicted values based on age and readiness.

Freshmen	Seniors
.514	.514
+ (.009 * 20.84)	+ (.009 * 27.51)
+ (.101 * -.0563)	+ (.101 * .2576)
.6959	.7876

Step C - calculate adjustment.
.7876 - .6959 = .0917

We first considered the freshmen, who had an average age of 20.84 and an average readiness score of -.0563. What would be the expected average topic 1 score of such a group, without regard to whether they were seniors or freshmen? Step B shows that it would be 69.59%. Likewise, the expected average score of the seniors, with an average age of 27.51 and an average readiness score of .2576, would be 78.76%. Step C shows that the difference between these – 9.17% – was what could be attributed to maturity and selection effects. In the fourth column of Table 3 this was subtracted from the gain to calculate the net gain, shown in the fifth column. Failure to make this adjustment would mean, in effect, that the program was taking credit for the fact that the students had grown older while in the program and that some poorly prepared students had dropped out, and of course credit should be taken only for what was actually taught to the students. For the sake of brevity, calculations for the adjustments for the other topics are not shown, but the procedure was the same in all cases.

The sixth column of Table 3 shows the standard deviation for the score on each topic, calculated using the whole data set including both seniors and freshmen, and the seventh column the net gain in standard deviations. This is necessary because a test resulting in a high standard deviation of scores, or one in which the difference between high performers and low performers is great, will also tend to show a greater difference between seniors and freshmen.

Discussion

Excluding Topic 5, the net gains for the content items shown in Table 3 ranged from 0.367 to 0.671

standard deviations. These results were encouraging because they exceeded Pascarella and Terenzini's estimate of 0.26 to 0.32 standard deviations for the net effect of attending college (2005, p. 71). The net gains for the two math items were 0.299 and 0.349 standard deviations. These results roughly match Pascarella and Terenzini's estimate of 0.29 to 0.32 standard deviations for the net effect of attending college on mathematical skills (2005, p. 71). Also, recall that the baseline in this study was the scores of freshmen *during* their introductory business course, after they had received some instruction. Had "pre-entry" scores been used, the net gains would likely have been even higher.

Topic 5, with a net gain of only .171, was a cause for concern. However, we should not immediately conclude that students are learning little in this area – it could be that the test questions did not effectively measure what they did learn. Also, note that the freshmen's raw score for topic 5 (column two of Table 3) was the highest of all the nine content and math items. It may be that the questions were too easy, making it hard to measure the freshman to senior gains. In any case, since assessment is an on-going process, this exercise should be repeated in coming semesters with different sets of questions, and more attention should be given to students' gains in topic 5 if disappointing results are repeatedly encountered.

However, the critical thinking (CT) scores were more troubling. The net gain for CT shown in Table 3 was only .093 standard deviations. Pascarella and Terenzini found few studies on gains in critical thinking or cognitive skills in general, but what they did find led them to an estimate of a gain of 0.50 standard deviations in critical thinking during college (2005, p. 205). This was a simple gain rather than a net gain, but Pascarella and Terenzini (2005) speculated that most of the gain was uniquely attributable to college, so that there should be little difference between the two. In any case, our finding of such a small net gain was a cause for concern, and this was especially true because the superior performance of the seniors on the multiple choice content items indicated that they did in fact have more background knowledge in business, which should be to their advantage in applying CT to the business situation used in our essay prompt (Bailin et al., 1999). Now, once again, we should not immediately conclude that students were not learning to think critically in the program – it is possible that they were, but that our method of evaluation did not capture these gains. Still, recall that the nine content and math items, for which we found substantial gains, were measured through multiple choice questions which simply require a student to recognize a correct answer when presented with one. Perhaps the program was primarily teaching these recognition skills, and neglecting more important CT skills.

These troubling results should provide a stimulus for the faculty to begin a process of improvement – the faculty may begin experimenting with new teaching techniques or curriculum in an effort to improve CT. Certainly, many authors have written about teaching techniques to improve CT. For example, Brookfield (2012) wrote that social learning can be used to teach CT. Students working in small groups to analyze scenarios – such as the one used in our essay prompt – are often surprised to find that their classmates interpret the scenario differently, and that this helps them to question their own reasoning and therefore to begin thinking critically. Pascarella and Terenzini (2005) describe a technique they called academic controversy, in which a group of four students is divided into two pairs and each pair is then assigned an opposite position on a controversial topic. Bonk and Smith (1998) described more than a dozen useful techniques including debates, mock trials, and case-based teaching, all of which may have advantages in teaching CT.

Limitations

In the following paragraphs we discuss five limitations of this study in what we consider to be ascending order of importance. First, as previously mentioned, we did not completely measure student gains within the program because the freshmen were not given the test at the beginning of their programs but rather at the end of their introductory class. Furthermore, giving the students a course embedded test, or one that counts as a portion of a grade, before they had received any instruction is probably impractical. Doing so would be quite intimidating to the students and very questionable from the perspective of fairness. However, we don't consider this to be a major flaw in our methodology because a test *near* the beginning of a program is a reasonable substitute for a test *at* the beginning of a program, and this results in a conservative estimate of student gains – we know that we underestimated rather than overestimated student gains within the program.

Second, we must address the issue of internally generated versus externally generated measures. Suskie (2009) referred to this issue as one of "local" versus "published" measures. We used a test that was written by teachers within the program rather than one written outside the program, such as a commercial test, and in our judgment it would be difficult if not impossible to find a commercial or published test flexible enough to be adapted to both a freshman and a senior level course. Again, however, we do not consider this to be a major flaw. We are aware that some external observers – including some administrators and even some accrediting agencies – are more impressed by externally generated measures. However, we consider curriculum-based measures a valid path of investigation. Such measures are touted in the field of education as an authentic form of measurement (Bradley, Danielson, & Hallahan, 2002).

And, as previously mentioned, internally generated tests have greater content validity than externally generated tests because they measure more specifically what students are expected to learn within a program, and therefore are more useful as a part of the four step assessment cycle mentioned in the introduction.

Third, in this study we made adjustments for age (the maturity effect) and readiness (the selection effect), but, even though we examined the available data for other material differences between the freshmen and the seniors, we can never be completely sure that these two variables were the *only* material differences. For example, there is considerable evidence that socioeconomic status affects academic achievement (Reardon, 2011). We attempted to examine this issue by asking students about their parents' education levels and their annual incomes, failed to find any differences between the freshmen and seniors, and therefore made no adjustment. Still, it is possible that there is some significant difference that we failed to find – perhaps there are socioeconomic differences that are not reflected in education or income. We do consider this to be a fundamental difficulty of our cross-sectional method of measuring learning, because while it will usually be possible to make adjustments for major effects such as maturity or selection, it will never be possible to ensure that *all* effects have been accounted for.

Fourth, our method is more easily applied to objective than to subjective measures of learning simply because it is easier to quantify objective measures. These include multiple choice tests, fill-in-the-blank tests for which no partial credit is given, or any other measure that can be scored without using professional judgment. Subjective measures, of course, are those requiring professional judgment to score (Suskie, 2009). In this study the objective measure was the multiple choice portion of the test, and applying our method to that portion required very little faculty labor except for making the test and data entry. The subjective portion was the critical thinking essay, and in order to apply our method we needed to quantify the students' performance on this portion. As described above, we first agreed upon a rubric, then scored the essays individually, and finally met to discuss and come to a consensus as to the scores. All this required a considerable amount of time and effort. Furthermore, we believe that quantifying subjective measures will inevitably be quite time consuming, and subjective measures are necessary to evaluate important skills such as critical thinking, creativity and problem solving. Since our objective is to propose a method for measuring learning under resource constraints, this is indeed a significant limitation.

Fifth, our method applies only to measurements of learning that can be quantified, and there are certainly some important educational outcomes, especially attitudes, that at least do not lend themselves to quantification. For example, most colleges of business seek to teach their students to be ethical, and most colleges of education seek to teach their students to value children and their differences. It is unlikely that these and other important attitudes will ever be reliably quantified, and so our method cannot be applied to them.

Conclusion

We hope that this paper has demonstrated a practical method of measuring learning, which is, of course, an important step in improving learning. Furthermore, this method, in which beginning and ending students are compared and regression techniques are used to make adjustments for important differences between the two groups, is applicable to any learning outcome so long as the outcome can be quantified; in other words, so long as a number can be assigned to each student's result. As we speculated in the previous section, it may be true that some important educational outcomes cannot be quantified, but regardless, there are many important outcomes that can be quantified. For example, Kraiger, Ford, and Salas (1993) and Kraiger (2002) described three kinds of outcomes. First, affective outcomes include confidence, attitudes and motivation, and these are of course the most difficult to quantify. But skill-based outcomes lend themselves more easily to quantification. For example, many institutions seek to teach and assess students' written or oral communication skills, and commonly assign grades to written work or presentations. The third category are cognitive outcomes, including declarative knowledge such as that which we measured through the content multiple choice questions in this study, and also cognitive strategies such as critical thinking, which we also measured. Most importantly, we believe that our method is both effective and within the means of nearly all institutions of higher education, including those with limited resources to devote to assessment.

References

ACT, Inc. (2010). *National collegiate retention and persistence to degree rates*. Retrieved from http://www.act.org/research/policymakers/pdf/retain_2010.pdf

Amaro-Jiménez, C., & Hungerford-Kresser, H. (2013). Implementing an additive, college access and readiness program for Latina/o high school students in the U.S. *Current Issues in Education, 16*(3), 1-12.

Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing critical thinking. *Journal of Curriculum Studies, 31*(3), 285-302.

Banta, T. W., Jones, E. A., & Black, K. E. (2009). *Designing effective assessment: Principles and profiles of good practice.* San Francisco: Jossey-Bass.

Bonk, C. J., & Smith, G. S. (1998). Alternative instructional strategies for creative and critical

thinking in the accounting curriculum. *Journal of Accounting Education, 16*(2), 261-293.

Bradley, R., Danielson, L., & Hallahan, D. P. (Eds.). (2002). *Identification of learning disabilities: Research to practice.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bresciani, M. J. (2006). *Outcomes-based academic and co-curricular program review: A compilation of institutional good practices.* Sterling, VA: Stylus Publishing, LLC.

Brookfield, S. D. (2012). *Teaching for critical thinking: Tools and techniques to help students question their assumptions.* San Francisco: Jossey-Bass.

Buskist, W., & Irons, J. G. (2008). Simple strategies for teaching your students to think critically. In D. S. Dunn, J. S. Halonen, & R. A. Smith (Eds.), *Teaching critical thinking in psychology: A handbook of best practices.* Sussex, UK: Blackwell Publishing Ltd.

Council for Higher Education Accreditation. (2014). Directory of CHEA-recognized organizations 2014-2015. Retrieved from http://www.chea.org/Directories/index.asp

Halpern, D. F. (2001). Assessing the effectiveness of critical thinking instruction. *The Journal of General Education, 50*(4), 270-286.

Halpern, D. F. (2003). *Thought and knowledge: An introduction to critical thinking* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kraiger, K. (2002). Decision-based evaluation. In K. Kraiger (Ed.), *Creating, implementing, and managing effective training and development: State-of-the-art lessons for practice* (pp. 331-376). San Francisco: Jossey-Bass.

Kraiger, K., Ford, J., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning to new methods of training evaluation. *Journal of Applied Psychology, 78*(2), 311-328.

Lovett, S., & Johnson, J. (2012). Measuring learning through cross sectional testing. *Journal of the Scholarship of Teaching and Learning, 12*(4), 43-57.

Meinert, D. (2013). Closing the Latino education gap. *HR Magazine, 58*(5), 28-33.

Oseguera, L., Locks, A. M., & Vega, I. I. (2009). Increasing Latina/o Students' Baccalaureate Attainment. *Journal of Hispanic Higher Education, 8*(1), 23-53.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research.* San Francisco: Jossey-Bass.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. J. Duncan & R. J. Murmane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 91-116). New York, NY: Russell Sage Foundation.

Rodgers, M., Grays, M. P., Fulcher, K. H., & Jurich, D. P. (2013). Improving academic program assessment: A mixed methods study. *Innovative Higher Education, 38*(5), 383-395.

Seifert, T. A., Pascarella, E. T., Erkel, S. I., & Goodman, K. M. (2010). The importance of longitudinal pretest-posttest designs in estimating college impact. *New Directions for Institutional Research, 2010*(S2), 5-16.

Suskie, L. (2009). *Assessing student learning: A common sense guide.* (2nd ed.) San Francisco: Jossey-Bass.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89-125.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). Chicago: University of Chicago Press.

U.S. Census Bureau. (2010). *Educational attainment in the United States: 2010 – Detailed tables.* Retrieved from https://www.census.gov/hhes/socdemo/education/data/cps/2010/tables.html

U.S. Department of Commerce. (2012). *State and county quickfacts.* Retrieved from http://quickfacts.census.gov/qfd/states/48/4810768.html

U.S. Department of Education. (2011). *Winning the future: Improving education for the Latino community*. Retrieved from http://www.whitehouse.gov/sites/default/files/rss_viewer/WinningTheFutureImprovingLatinoEducation.pdf .

U.S. Department of Labor. (2011). *The Hispanic labor force*. Retrieved from http://www.dol.gov/_sec/media/reports/hispaniclaborforce.pdf.

Walvoord, B. E. (2010). *Assessment clear and simple: A practical guide for institutions, departments and general education* (2nd ed.). San Francisco: Jossey-Bass.

Zumeta, W. M. (2011). What does it mean to be accountable? Dimensions and implications of higher education's public accountability. *The Review of Higher Education, 35*(1), 131-148.

Article Citation
Lovett, S., & Curtis, M. G. (2015). Assessment under resource constraints. *Current Issues in Education*, *18*(1). Retrieved from http://cie.asu.edu/ojs/index.php/cieatasu/article/view/1357

Author Notes
Steve Lovett
The University of Texas at Brownsville - School of Business
One West University Blvd., Brownsville, Texas 78520
steve.lovett@utb.edu

Dr. Steve Lovett is an Associate Professor in the Department of Management and Marketing at the University of Texas at Brownsville. His research has focused on the assessment cycle and the measurement of learning.

Mary G. Curtis
The University of Texas at Brownsville - College of Education
One West University Blvd., Brownsville, Texas 78520
mary.curtis@utb.edu

Dr. Mary Curtis is an Associate Professor of Special Education in the Educational Psychology and Leadership Studies at the University of Texas at Brownsville. Her research interests focus on the assessment of language acquisition, language differences, and language disorders.

Current Issues in Education
Mary Lou Fulton Teachers College • Arizona State University
PO Box 37100, Phoenix, AZ 85069, USA

Manuscript received: 01/15/2014
Revisions received: 09/04/2014
Accepted: 09/26/2014

Current Issues in Education

Mary Lou Fulton Teachers College • Arizona State University
PO Box 37100, Phoenix, AZ 85069, USA

Volume 18, Number 1 March 22, 2015 ISSN 1099-839X