Issues in institutional benchmarking of student learning outcomes using case examples

Thomas P. Judd
United States Military Academy

Christopher Pondish
City University of New York

Charles Secolsky
Savoy, Illinois

ABSTRACT

Benchmarking is a process that can take place at both the inter-institutional and intra-institutional level. This paper focuses on benchmarking intra-institutional student learning outcomes using case examples. The findings of the study illustrate the point that when the outcomes statements associated with the mission of the institution are criterion-oriented and not comparative, then benchmarking can take place with respect to institutional standards or competencies. Another form of intra-institutional benchmarking, known as normative assessment, can occur when students from different majors are compared with respect to common core skill areas. Both types of the intra-institutional *context*s for benchmarking, criterion-oriented or normative, depend on the mission-related institutional standards of performance. Issues identified relate to the potential for inappropriate or invalid inferences being made from outcomes assessment results using rubrics and baselines due to a lack of validity, primarily as a result of what might still be proper statistical applications. Benchmarking is important to educational decision-making processes. Yet, more thought is needed on how human judgment surrounding the benchmarking process influences the validity of curricular decisions and relationships to student learning outcomes. Offered is an in-depth understanding of benchmarking types and how to further their uses with student learning outcomes.

Keywords: Benchmarking, Student learning outcomes, Baselines, Intra-institutional and Inter-institutional benchmarking, Context, Standards

.

INTRODUCTION

For the past several years, benchmarking has become an important part of the educational decision-making process. Yet little thought is given to how judgments surrounding the benchmarking process influence the validity of curricular decisions and their relationship to student learning outcomes. Benchmarks are naturally occurring phenomena that are a subset of standard-setting methodology and as such may be overly used without careful consideration of the role that human judgment plays in selecting the benchmark. The authors believe there is an inverse relationship between the amount of interpretability in the benchmarking criteria and the validity of the benchmark in accurately reflecting student learning outcomes. They recommend that the use of benchmarking should be restricted to circumstances where the focus is not on politically redistributing or redefining rewards in education but only for improving student learning outcomes, given the thoughtful and relatively unbiased construction of the benchmark. To this end, this paper examines how some issues associated with the benchmarking process can be addressed.

Benchmarking is a process that can take place at both the inter-institutional and intra-institutional level. This paper focuses on benchmarking intra-institutional student learning outcomes using case examples. The findings of the study illustrate the point that when the outcomes statements associated with the mission of the institution are criterion-oriented and not comparative, then benchmarking can take place with respect to institutional standards or competencies. Another form of intra-institutional benchmarking, known as normative assessment, can occur when students from different majors are compared with respect to common core skill areas. Both types of the intra-institutional *context*s for benchmarking, criterion-oriented or normative, depend on the mission-related institutional standards of performance. Issues identified relate to the potential for inappropriate or invalid inferences being made from outcomes assessment results using rubrics and baselines due to a lack of validity, primarily as a result of what might still be proper statistical applications (see Judd & Keith, 2012).

This study of internal benchmarking of student learning outcomes uses comparisons of the same or similar learning outcomes for individual courses within an institution and for the same courses over time for the purposes of formative assessment or improvement. According to Upcraft and Schuh (1996) and Seybert, Weed and Bers (2012), there are three types of benchmarking: internal, generic and comparative. Spendolini (1992) describes internal benchmarking as the process of comparing practices within an institution. Intra-institutional benchmarking, by this definition, is internal benchmarking. Existing interdepartmental cultural differences make intra-institutional benchmarking of student learning outcomes at a given institution difficult to accomplish. Similarly, inter-institutional benchmarking may be even more difficult to achieve due to institutional as well as a departmental contextual nexus of factors. If benchmarking is to have greater potential for validation within and across institutions, then a greater degree of standardization is desirable. In the case of student learning outcomes, standardization can refer to the curriculum, intended student learning outcomes, evaluation methods, assessment instruments, mode of instruction (online vs. face-to-face), and/or testing. Intra-institutional benchmarking involves making comparisons between units within the same institution. Earlier work on norm-referenced testing could be considered a backdrop for the development and conceptual framework of the benchmarking movement (Upcraft & Schuh, 1996). Background on intra-institutional benchmarking with respect to

learning outcomes emerges from the criterion-referenced literature in which each item of a test or a task is defined by some domain of interest and success on that domain over time. Optimally, the item or task being benchmarked would be representative of some set of admissible observations. In cases such as the use of rubrics and their inherent criteria, the variance of the scores associated with the observations can be parsed into three categories: variance attributable to criteria; variance attributable to raters; and, variance attributable to their interaction. The variance attributable to raters and the interaction of criteria and raters provides fodder for arguments against the use of baselines when evaluating student learning outcomes.

The Difference between a Benchmark and a Standard

For rubric based benchmarks, the disaggregation of data (the individual criteria for each rubric) fosters additional utility than that of the overall rubric score which is an effort to make judgments about success. Within that context, benchmarks can be, more often than not, naturally occurring (e.g., comparison to best in class) and consequently can vary by department and/or institution. As noted above, the variance inherent in rubric evaluation can call into question the inter-rater reliability of the benchmark, although intra-institutional use can often yield consistent results. A standard, on the other hand, is based more so on human judgment and the harnessing of those judgments to arrive at a cutscore or minimum level of acceptable performance. The fact is that a score of 65 is passing is based on judgment which has been accumulated over a number of years. Standards can be adjusted through validation by a consensus of external expertise, as determined by disciplines with national or regional accrediting agencies, or, in the case of university systems, a common core of system-wide standards (Judd & Keith, 2012). A standard can also be validated by examining scores on a commonly accepted external criterion, hence the rise of standardized testing. Success based on such criteria would indicate that students have met or exceeded the standard set and are predominantly scoring successfully on the criterion for which the standard was based.

Once benchmarks have been identified, the focus can turn to what legitimate uses can be made of intra-institutional benchmarking data on student learning outcomes. Before this question can adequately be answered, there are a number of obstacles that need to be overcome, especially if decisions made based on benchmarking data can be used to effectuate change. Foremost, faculty development initiatives are needed to overcome resistance (possibly related to issues of validity and/or academic freedom) to benchmarking. Even if faculty has accepted that assessment information legitimizes the need for improvement, changing curriculum based on benchmarking can be a delicate balancing act. Resistance can also result from politically charged comparisons among the institutional departments or disciplines. Outcomes assessment has traditionally meant closing the loop after an intervention has taken place. Typically, the intervention can be theoretical and curricular validity may be unknown or lacking since student ability is often not controlled. The same can be true for differences between classes of the same course with different instructors or classes from different institutions with different learning environments, all of which add to the validity of benchmarking debate (see Seybert, Weed and Bers, 2012).

Types of Intra-Institutional Benchmarking for Student Learning Outcomes

Standards-based benchmarking seeks to determine how good student performance needs to be to meet the learning outcomes (see Stake, 2004). A second type of benchmarking establishes a criterion of performance growth or progress over time using baselines (see American Society for Quality 2011). A third type of benchmarking can take place with respect to indirect measures of student learning outcomes such as measured by the constructs from the National Survey of Student Engagement (NSSE). For the first type of benchmarking, the answer to how much is good enough requires that a point on the skill or ability continuum is determined that represents adequate or expert attainment for the skill or ability one is assessing. For intra-institutional benchmarking of student learning outcomes, defining such a benchmark requires some form of standard-setting. The field of standard-setting in educational measurement is based on judgment and in some rare instances can be empirical, employing different methodologies to accomplish this purpose (see Pitoniak & Morgan, 2012).

METHODOLOGY USED FOR UNCOVERING IMPORTANT ISSUES

Three case examples were used to illuminate issues surrounding the three types of benchmarks in an intra-institutional setting. The first case is a comparison of rubric scores for capstone courses in Graphic Design and Photography offered by the same Graphic Design Department at a community college. The same rubric was used for both courses. The second case example shows the progress of students over a course sequence in mathematics. It exemplifies the creation of a trend and establishes a baseline by providing pass rates of students starting in a developmental Intermediate Algebra course through a course in College Algebra through Pre- Calculus through Calculus. The second case demonstrates the advantages and disadvantages of using pass rate trends as benchmarks. Finally, benchmarking is discussed using constructs from the NSSE data on a large sample as indirect measures, and the potential for misleading interpretations of intra-institutional comparisons. See Table 1, Appendix.

Example 1:Benchmarking standards across courses

The rubric used in Table 1 has four criteria or dimensions: Technique, Design, Creativity and Concept, and Presentation. Discretized continuous point allocations with descriptions appear in each of the 16 cells of the rubric. For both the Graphic Design and Photography courses, there were the same four judges or raters. Averages in the form of means were computed for each criterion across the judges.

Example 2: Benchmarking growth

In a four-course mathematics sequence the same six questions representing different domains of skill or ability were embedded in the final exams at the end of the semester, but did not count in students' grades. The problem addressed by embedding the questions was to what extent should students as an aggregate answer each of the six questions correctly as they progress through the sequence? Three consecutive semesters of data: spring 2009, fall 2009, and spring 2010 were used in the math case example. Analyses produced results for each item by course and as a subtest of all six items by course.

Example 3: Benchmarking indirect measures

Issues in intra-institutional benchmarking are identified and the difficulties encountered by comparing NSSE benchmarks across college departments are discussed. With very large numbers of cases at some institutions and the larger number of cases for the peer groups, differences using t-tests are often significant, but practical significance may be minimal. For this purpose, effect sizes, which are unaffected by large sample sizes, are employed with NSSE data.

RESULT

Rubric scores for the Graphic Design and Photography courses are presented in Table 2 and Table 3, respectively. The rubric scoring is typical of rubric scoring associated with portfolio assessments for determining impact of instruction on a particular curriculum. For each criterion, as is the case for Graphic Design and Photography, limited statistical comparisons are often made. The authors surmise that this is the case because meaning is imputed in the rubric cells and an attempt is made by scorers of rubrics to keep statistical analyses simple, reflecting at most the mean or average criterion score, and at times providing data on inter-rater-reliability or agreement apart from the central analysis of the rubric scoring. See Tables 2 and 3, Appendix.

Common standards across different courses

There were 17 work products for Graphic Design and 13 work products for Photography. An inspection of the rubric scoring for Graphic Design shows mean ratings ranging 91.5 to 97.32 for the four criteria with Presentation having the highest mean. In Table 3 (to a large extent the same criteria) means ranged from 77.8 to 80.6. The same four judges rated the work assignments higher for Graphic Design on average than the student work products for Photography. But does this inference tell the whole story of the benchmarking of these two courses? **The** answer is an emphatic "no." First there are students who have outlier performances that lower the mean score considerably. Student #8 for Photography – a score of 53 on Technique, or the low scores for students #8 and #9 for Creativity is an extreme score that unduly influences the mean with so few scores in the calculation.

Obviously, the basic statistical concept of a standard deviation of rubric scores for each criterion would potentially provide some pivotal information if these portfolio assessments were used to assess outcomes. In fact, a generalizabilty theory study would enable the parsing out of variance attributable to criterion effect, rater effect , and their interaction term (Webb, Shavelson & Steedle, 2012) This would help to identify instances where outcomes assessment interventions could be more effective than when the rater effect is carrying or masking the differences between criteria (see Secolsky & Judd, 2011). See Tables 4 - 11, Appendix.

Mathematics progress

The trends for most of the items from the analysis of the math course products indicated that higher percentages of students responded with the correct answer as the course material became more advanced. This result is what was expected. However, for item 2, a smaller percentage of students in M131 (Calculus) answered the item correctly in comparison to those

in M123 (pre-calculus). By benchmarking courses against each other, it was possible to identify where students who were learning more advanced material in mathematics demonstrated less of an ability to respond correctly than students from a less advanced course. The same was true for item 5 on the perimeter question. Nevertheless, there was a clear progression of percent of students responding correctly to the items as the level of the course offering became more advanced. The identification of these two items provides an example of how benchmarking activities help make the full circle from assessment to planning. The next step in the process is to undertake the development of a plan to improve performance in the skills required by items 2 and 5, implement that plan and repeat benchmark assessment measures.

By looking at the percent of students responding to the non-correct distracters, benchmark trends like the ones in Tables 4-10 may help to identify differences in how items were conceptualized by students. For item 6, upon choosing the correct equation for the graph, the group in M016 (Intermediate Algebra) had only a 33.7% pass percentage for this item as compared to 92.3% for M123 (Pre-Calculus) and 93.9% (for Calculus), had a 7.2% responding to incorrect distracter (b).

NSSE Benchmarks

Intra-institutional benchmarking using NSSE data, while providing very valuable comparative information between departments for an institution as well as student characteristics, can at times be problematic for two reasons. First, there is no absolute standard – department means are compared to one another with respect to NSSE questions. While one department can exceed the mean for another department on a given question, Stake's(2004) point of how good should the outcomes be introduces the idea of the relative nature of benchmarking applied to a particular context. Coupled with this relative nature is the charged political comparisons that may develop as a result of intra-institutional benchmarking. Nonetheless, other types of intra-institutional benchmarking could be performed such as those between freshman and fourth year students, males and females and athletes and non-athletes on a given campus although, comparisons based on such antecedent characteristics such as gender carry their own wealth of political dynamite.

Table 12 and Table 13 show differences between athletes and others on NSSE benchmarks for males and females and first and fourth year students on select NSSE constructs representing Academic Challenge, Active and Collaborative Learning, Student-Faculty Interaction, Enriching Educational Experiences and Supportive Campus.. Six of the comparisons in the two tables produced statistical significance via an independent samples t-test (scores in bold), yet these potentially charged differences can be misleading.

Significance testing with the independent samples t-test is a standard method for comparing the means of each department, either to other departments or to the overall mean of the college. Significance testing, however, is influenced by the sample size. With larger sample sizes, statistical significance can be found with relatively small differences in means, which can make interpretations of meaningful differences challenging. The effect size statistic is independent of sample size, and can be more helpful in identifying differences that have practical significance. It is useful to note the relationship between sample size and effect size because it plays a role when interpreting results across classes, departments, institutions and systems. Effect sizes can be calculated using Cohen's d statistic, which is the difference between the means divided by the pooled standard deviations. Effect sizes between .2 and .5

are considered small, between .5 and .8 are considered medium and effect sizes of .8 or over are considered large effects. Effect sizes can be calculated for each comparison of means where statistical significance is found using the independent samples t-test.

The data in Tables 12 and 13 represent a sample of 1,734 students, qualifying it as a large sample size, as described above, with the attendant challenge of interpreting practical significance in a manner that does not distort the findings. Effect size calculations for the six statistically significant comparisons were all less than the threshold of .2 for small effects, rendering the differences between the groups as relatively meaningless, and certainly not worthy of extended institutional action. They are sufficiently noteworthy to monitor over the course of several years. See Tables 12 and 13, Appendix.

DISCUSSION/CONCLUSIONS

The politically charged issue of comparing departments intra-institutionally may be somewhat ameliorated by providing each department their own data in comparison to the overall mean, allowing them to freely share and compare with other departments as they wish. Some important questions to be considered as this study concludes include: Is it possible to equate outcomes measures? Can the instruction, task, test, and item design be made comparable? Can item response theory be used to allow for sample-free ability estimation so that students' scores can be compared? All these hypothetical questions need to be considered as viable avenues for the future as the link between student learning outcomes and budgetary constraints take on greater importance. Bench markers should pay particular attention to the use of benchmarks when dealing with nominal variables and other qualitative measures. Bearing in mind that the operationalization of multi-element, judgmental conceptual measurements is always a highly subjective if not questionable practice, the authors recommend benchmarking such measurements is, perhaps, not as viable an option as the current acceptable practices may suggest. Additional research and discussion of this point is merited.

REFERENCES

American Society for Quality (2011). *Criteria for performance excellence 2011-2012*, Gaithersberg, MD, National Institute of Standards and Technology.

Judd, T. and Keith, B. (2012). *Student learning outcomes at the program and institutional levels*. In C. Secolsky & D. B. Denison (Eds.) *Handbook on measurement, assessment, and evaluation in higher education* (pp. 31-46) New York: Routledge.

Pitoniak, M, & Morgan, D. (2012). *Setting and validating cutscores for tests*. In C. Secolsky & D. B. Denison. (Eds.) *Handbook on measurement, assessment, and evaluation in higher education* (pp. 343-366) New York: Routledge.

Secolsky, C. & T. Judd (2011). *Rubrics and outcomes assessment: A generalizability theory approach.* Paper presented at the Annual Forum of the Association for Institutional Research, Toronto, Canada.

Seybert, J., Weed, E. & Bers, T. (2012). *Benchmarking in higher education*. In C. Secolsky & D. B. Denison (Ed.) *Handbook on measurement, assessment, and evaluation in higher education.*(pp. 100-122) New York, Routledge.

Speldoni, M. J. (1992). *The benchmarking book*. New York: American Management . Association.

Stake, R.E (2004). *Standards-based and responsive evaluation*. Thousand Oaks, CA. Sage.
Upcraft, M.L. & Schuh, J. H. (1996). *Assessment in student affairs: A guide for practitioners*: San Francisco: Jossey-Bass Higher and Adult Education.
Webb, N., Shavelson, R. & Steedle, J. (2012). *Generalizability theory in assessment contexts*. In C. Secolsky & D. B. Denison (Ed.). *Handbook on measurement, assessment, and evaluation in higher education* (pp. 132-149) New York, Routledge

APPENDIX

.

Table 1: OUTCOMES ASSESSMENT RUBRIC FORM

Component Possible 100 Points	**Outstanding 25 Points**	**Highly Successful 20 Points**	**Successful 15 Points**	**Not Yet Successful 10 Points**
Technique	Very good understanding of different media and their uses. Work exhibits mastery of visual arts techniques.	Good understanding of different media and their uses. Work exhibits good control of visual arts techniques.	Solid understanding of different media and their uses is not very broad. Work exhibits competence of visual arts techniques.	Understanding of different media and their uses is not evident. Work exhibits limited mastery of visual arts techniques.
Design	Very good understanding of the elements good design and composition and uses these, skillfully and effectively to communicate ideas.	Good understanding of the elements good design and composition and uses these very well to communicate ideas in most instances.	Solid understanding of the elements good design and composition. Communication established but unintended.	Understanding of the elements good design and composition is not evident. Communication skills are poor.
Creativity and Concept	Work is unique and presents an original, interesting and clear conceptualization of an idea.	Work is mostly unique and presents a largely original, interesting and clear conceptualization of an idea.	Work contains unique and derivative elements and presents a partially original, interesting and clear conceptualization of an idea.	Work is derivative, uninteresting and lacks clarity.
Presentation	Work exhibits mastery of skills and materials without error.	Work exhibits appropriate use of skills and materials without significant errors.	Work exhibits a rough approximation of what is appropriate, includes a few errors.	Work exhibits critical errors in the use of materials or skills specific to the task.

Table 2: Ratings from Graphic Design Rubric

Student	Technique	Design	Creativity and Concept	Presentation	Total Points
1	95	90	85	100	370
2	100	90	95	100	385
3	90	90	90	100	370
4	90	85	85	100	360
5	87	95	89	95	366
6	80	85	85	95	345
7	94	100	100	100	394
8	94	95	95	100	384
9	94	95	95	95	379
10	99	100	100	95	394
11	89	90	90	95	364
12	100	95	100	100	395
13	90	95	95	95	375
14	89	85	90	95	359
15	95	100	99	100	394
16	90	85	90	95	360
17	80	90	85	95	350
18					
Average	1556/4=389 389/17=22.88 22.88x4=91.5 91%	1564/4=391 391/17=23 23x4=92 92%	1568/4=392 392/17=23 23 x4=92 92%	1658/4=413.75 413.75/17=24.33 24.33x4=97.32 97%	

Table 3: Ratings from Photography Rubric

Student	Technique (Avg.)	Design	Creativity	Presentation	Total Points (# of reviews)
1	72 (18)	70 (17.5)	71 (17.75)	70 (17.5)	283 (4) = 70.75%
2	102 (20.4)	103 (20.6)	106 (26.5)	102 (20.4)	413 (5) = 82.6%
3	91 (22.75)	89 (22.25)	82 (20.5)	83 (20.75)	345 (4) = 86.25%
4	85 (21.25)	89 (22.25)	84 (21)	85 (21.25)	343 (4) = 85.75%
5	93 (18.6)	98 (19.6)	97 (19.4)	100 (20)	388 (5) = 77.60%
6	90 (22.5)	90 (22.5)	88 (22)	96 (24)	364 (4) = 91.00%
7	87 (21.75)	84 (21)	75 (18.75)	81 (20.25)	327 (4) = 81.75%
8	53 (13.25)	63 (15.75)	63 (15.75)	65 (16.25)	244 (4) = 61.00%
9	87 (21.75)	78 (19.5)	67 (16.75)	93 (23.25)	325 (4) = 81.25%
10	76 (19)	84 (21)	86 (21.5)	78 (19.5)	324 (4) = 81.00%
11	76 (19)	72 (18)	65 (16.25)	80 (20)	293 (4) = 73.25%
12	92 (18.4)	102 (20.4)	102 (20.4)	100 (20)	396 (5) = 79.20%
13	65 (16.25)	77 (19.25)	79 (19.75)	75 (18.75)	296 (4) = 74.00%
Total Points	252.9 / 13 = 19.45	259.6 / 13 = 19.96	256.30 / 13 = 19.71	261.9 / 13 = 20.15	4341 / 55 = 79%
Average	19.45 / 25 = **77.8%**	19.96 / 25 = **79.84%**	19.71 / 25 = **78.84%**	20.15 / 25 = **80.6%**	= **79.27%**

Table 4

1. Solve the equation $\frac{2}{5x}+3=6$ Answer $\frac{2}{15}$

	M016	M110	M123	M131
a) $\frac{2}{45}$	5.8%	4.1%	1.9%	0.0%
b) $\frac{2}{15}$	47.9%	61.7%	88.7%	97.0%
c) $\frac{45}{2}$	5.8%	2.1%	0.0%	3.0%
d) $\frac{15}{2}$	21.8%	16.2%	7.5%	0.0%
e) None of the above	18.8%	15.9%	1.9%	0.0%

Table 5

2. 117 is 65% of what number? Answer 180

	M016	M110	M123	M131
Spring 2009	66.7%	75.4%	89.8%	91.7%
Fall 2009	34.4%	72.4%	88.2%	84.1%
Spring 2010	59.8%	74.2%	83.5%	88.1%

Table 6

3. $(5x-2)^2 =$	Answer	$25x^2-20x+4$		
	M016	M110	M123	M131
a) $25x^2-4$	8.2%	4.1%	0.0%	0.0%
b) $25x^2-20x-4$	4.2%	10.0%	13.2%	12.1%
c) $25x^2+4$	6.6%	5.6%	0.0%	0.0%
d) $25x^2-20x+4$	29.7%	79.0%	86.8%	87.9%
e) None of the above	1.3%	1.3%	0.0%	0.0%

Table 7

4. Solve the equation $x^2-3x=28$	Answer 7 or -4			
	M016	M110	M123	M131
a) -7 or 4	8.1%	12.9%	1.9%	9.1%
b) 7 or -4	29.3%	73.3%	82.7%	90.0%
c) -7 or -4	3.8%	4.4%	1.9%	0.0%
d) 7 or 4	4.1%	2.1%	5.8%	0.0%
e) None of the above	4.7%	7.5%	7.7%	0.0%

Table 8

5. The perimeter P of a rectangular yard is 330 feet. The length is 75 feet more than twice the width. Find the width. $(P = 2l + 2w)$

Answer w = 30 feet

	M016	M110	M123	M131
Spring 2009	34.6%	57.0%	72.7%	78.3%
Fall 2009	20.9%	55.7%	76.9%	72.9%
Spring 2010	34.9%	58.7%	71.1%	80.6%

Table 9

2. 117 is 65% of what number? Answer 180

	M016	M110	M123	M131
Spring 2009	66.7%	75.4%	89.8%	91.7%
Fall 2009	34.4%	72.4%	88.2%	84.1%
Spring 2010	59.8%	74.2%	83.5%	88.1%

Table 10

5. The perimeter P of a rectangular yard is 330 feet. The length is 75 feet more than twice the width. Find the width. $(P = 2l + 2w)$

Answer w = 30 feet

	M016	M110	M123	M131
Spring 2009	34.6%	57.0%	72.7%	78.3%
Fall 2009	20.9%	55.7%	76.9%	72.9%
Spring 2010	34.9%	58.7%	71.1%	80.6%

Table 11

Percentage of Students with Correct Answers

Semester	M016	M110	M123	M131
Spring 2009	52.6%	70.2%	87.5%	89.2%
Fall 2009	57.3%	69.7%	85.9%	87.7%
Spring 2010	50.9%	70.0%	80.5%	86.1%

Table 12: Means of Athletes* and Other Students on NSSE Benchmarks

	N	Combined	First Year	Seniors
Academic Challenge	Athlete	61.9	61.8	62.2
	Others	61.9	61.6	62.2
Active and Collaborative Learning	Athletes	54.0	51.4	59.4
	Others	54.2	51.0	57.4
Student-Faculty Interaction	Athletes	49.6	**45.3**	58.4
	Others	48.8	**41.6**	55.9
Enriching Educational Experiences	Athletes	39.3	31.2	**55.4**
	Others	41.0	30.3	**51.5**
Supportive Campus Environment	Athletes	**64.2**	64.8	**63.1**
	Others	**62.0**	64.7	**59.2**

Statistically significant differences between athletes and others means are in **bold.*** For the NSSE analysis, *athletes* are defined as those students who report yes to the item *Are you a student-athlete on a team sponsored by your institution's athletics department***.**

Table 13: Means of Male and Female Students on NSSE Benchmarks

	N	Combined	First Year	Seniors
Academic Challenge	Male	61.8	61.6	62.1
	Female	62.2	62.0	62.5
Active and Collaborative Learning	Male	**54.4**	51.3	**58.7**
	Female	**52.2**	50.3	**54.8**
Student-Faculty Interaction	Male	49.4	43.6	57.3
	Female	47.6	43.2	53.5
Enriching Educational Experiences	Male	39.9	30.4	52.8
	Female	41.5	32.3	53.8
Supportive Campus Environment	Male	63.3	64.8	61.1
	Female	62.0	64.6	58.5

Statistically significant differences between male and female means are in **bold.**